Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hallmark pathway scores and prognosis in lower-grade glioma: validation in two cohorts against random gene sets

Aya Darwish

Ayavet27@hotmail.com

Pathfinder Epidemiology Academy

Ahmed Kassab

Benha University

Moustafa Sherif

Benha University

Research Article

Keywords: Diffuse glioma, survival prediction, Hallmark pathways, random gene sets, external validation.

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11677226/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hallmark pathway scores and prognosis in lower-grade glioma: validation in two cohorts against random gene sets

Aya Darwish [1], Ahmed Kassab [2], Moustafa Sherif [1, 2]*

[1] Pathfinder Epidemiology Academy, Al-Obour , Egypt; Ayavet27@hotmail.com
[2] Faculty of Veterinary Medicine, Benha University, Toukh 13736, Egypt; ahmed.kassab@fvtm.bu.edu.eg, mustafa.shref@fvtm.bu.edu.eg

* Corresponding author: Aya Darwish
Address: Pathfinder Epidemiology Academy, Al-Obour , Egypt
Email: Ayavet27@hotmail.com
Tel: +2 01006824606
ORCID: 0009-0009-7251-5525

Abstract

Background: Hallmark pathway scores may capture prognostic biology in lower-grade glioma, but their added value beyond clinical and molecular factors remains uncertain. **Methods**: We analyzed TCGA-LGG (352 patients, 87 deaths) and CGGA (561 patients, 329 deaths). Clinical Cox models were compared with models adding Hallmark pathway scores (eight selected pathways, all 50, or five components) through cross-validation and bidirectional external validation. Random expression-matched gene sets and whole-transcriptome components served as benchmarks. **Results**: The clinical model achieved C indices of 0.85 in TCGA cross-validation, 0.75 in forward validation, and 0.82 in reverse validation. Pathway scores changed the C index by −0.017 to +0.015 across four primary evaluations, and no change remained significant after Holm adjustment. Five pathway components added information beyond clinical predictors in pooled analyses ($\chi^2 = 25.1$, 5 df, P = 0.00013), while random sets and whole-transcriptome components provided comparable or greater information. Sixteen pathways were nominally more prognostic than matched random sets in both cohorts. Eight-pathway scores selected in one cohort were associated with survival in the other cohort without a consistent gain in C index. **Conclusions**: Hallmark pathway scores did not consistently improve discrimination beyond clinical and molecular factors. Some pathways were more prognostic than matched random sets in both cohorts, and random-set benchmarks help separate pathway specific from generic transcriptomic information.

Keywords

Diffuse glioma, survival prediction, Hallmark pathways, random gene sets, external validation.

Background

Diffuse lower-grade gliomas, WHO grade II and III tumors in adults, vary substantially in survival. An integrative analysis of the TCGA cohort showed that IDH mutation status and 1p/19q codeletion distinguish three groups with different clinical outcomes (Brat et al., 2015; Eckel-Passow et al., 2015). These markers are incorporated into the WHO classification of adult diffuse gliomas (Louis et al., 2021). Age, tumor grade, and molecular subtype remain central factors in prognosis.

Research has examined whether transcriptomic and multiomics signatures improve prognostic prediction beyond these factors. A multiomics model for lower grade glioma reported area under the curve (AUC) values of 0.817

in the training set and 0.819 in the test set (Du et al., 2024). Such results do not establish the added value of a molecular signature unless it is compared with a clinical model evaluated under the same validation scheme. In breast cancer, many randomly selected gene expression signatures were associated with outcome (Venet et al., 2011). This finding illustrates why an association with survival alone does not establish biological specificity. Gene set methods also test different null hypotheses. Some approaches treat genes as independent sampling units and can produce P values that are too small when gene dependence is ignored (Goeman & Bühlmann, 2007). External validation is important for assessing discrimination and calibration in independent data (Ramspek et al., 2021; Royston & Altman, 2013).

The Cancer Genome Atlas (TCGA) and the Chinese Glioma Genome Atlas (CGGA) provide public glioma cohorts with clinical annotations and RNA sequencing data (Brat et al., 2015; Zhao et al., 2021). We used these cohorts to test whether curated Hallmark pathway scores improve survival prediction beyond age, WHO grade, and molecular subtype, and whether any improvement is specific to biological pathways. We assessed changes in discrimination and prediction error using models trained in one cohort and applied without refitting to the other. We then reversed the training and testing cohorts. We also compared the additional likelihood information from pathway scores with that from random gene sets matched for size and components derived from the whole transcriptome. Finally, we identified pathways that were more prognostic than random sets in both cohorts and tested whether a score based on these pathways predicted survival in the other cohort.

Methods

This study followed the TRIPOD+AI reporting guideline (Collins et al., 2024). All analyses used publicly available, deidentified data.

TCGA lower grade glioma data were obtained from the Genomic Data Commons (GDC). Gene expression for protein coding genes was taken from the harmonized STAR workflow as transcripts per million (TPM). Duplicate files were checked using MD5 checksums, and five recurrent tumor files were replaced with primary tumor files. Overall survival and progression free interval were obtained from the TCGA Clinical Data Resource (TCGA-CDR, (Liu et al., 2018)). Follow up times were converted from days to months by dividing by 30.44. Age at diagnosis came from the GDC, with the TCGA-CDR value as fallback. WHO grade came from the TCGA-CDR histological grade, with the GDC grade as fallback.

IDH status and 1p/19q codeletion were obtained from the cBioPortal study lgg_tcga_pan_can_atlas_2018. Tumors with wild type IDH were classified as IDH wild type regardless of codeletion status. IDH mutant tumors were classified as codeleted or noncodeleted, and those without a codeletion label were excluded. Somatic mutations were obtained from GDC masked ensemble mutation annotation format files for primary tumors. Nonsynonymous variants included missense, nonsense, frameshift, in frame, splice site, translation start site, and nonstop variants. Mutation burden was the natural logarithm of one plus the number of nonsynonymous variants per patient. Genes mutated in at least 3% of covered cases were represented by binary indicators. TCGA eligibility required expression data, positive follow up time, resolved IDH and 1p/19q status, and an available mutation file. One patient was excluded only because the mutation file was missing. The same eligible cohort was used for all TCGA models.

CGGA data (Zhao et al., 2021) comprised the mRNAseq_693 and mRNAseq_325 batches. FPKM expression values were converted to TPM, and gene symbols were harmonized with TCGA. Clinical variables included age, WHO grade, IDH status, 1p/19q codeletion, survival time, and MGMT promoter methylation where available. Eligible patients had a primary tumor, positive survival time, recorded age, and resolved IDH and 1p/19q status. Follow up times were converted to months using the same procedure as in TCGA. Batch was retained for stratification in association analyses. TCGA expression and mutation files were downloaded from the GDC between 2 September and 2 October 2026 (Data Release 46.0), the TCGA-CDR workbook on 29 September 2026, and the CGGA files on 8 and 9 September 2026. The cBioPortal study had a data import date of 3 June 2026.

Expression values were log base 2 transformed after adding one to each TPM value, and analyses used the 15,622 genes shared by both cohorts. The 50 Hallmark gene sets were obtained from the Enrichr MSigDB_Hallmark_2020 library using GSEApy (Fang et al., 2023; Liberzon et al., 2015). Each set retained at least three shared genes. Pathway scores were calculated as the mean standardized expression of the genes in each set. For transferred models, gene means and standard deviations were estimated in the training cohort and applied unchanged to the test cohort. Sensitivity analyses used test cohort and batch specific standardization or single sample gene set enrichment analysis (ssGSEA; (Barbie et al., 2009)).

Overall survival was the primary outcome, with TCGA progression free interval used for sensitivity analysis. The reference Cox model included age, ordinal WHO grade, and indicators for IDH mutant codeleted and IDH mutant noncodeleted tumors, with IDH wild type as the reference group. Covariates were standardized using the training data. Three pathway models added the eight pathways most strongly associated with survival after adjustment for clinical variables, all 50 pathway scores, or the first five principal components of the pathway scores. Pathways in the first model were selected within each training set. Ridge penalization was applied to added terms, while clinical terms remained unpenalized. The penalty was selected from 0.001 to 30 using inner fivefold cross validation of the partial log likelihood.

For the eight pathway model, selection occurred once before penalty tuning, so patients in inner validation folds contributed to selection. Outer test folds and external cohorts were not used for selection or tuning. A sensitivity analysis repeated pathway selection within each inner split and extended the penalty grid to 3,000. All preprocessing, including expression standardization, pathway scaling, principal component loadings, and mutation feature scaling, was estimated in the training data and applied unchanged to test data. Random survival forests (Ishwaran et al., 2008) with 200 trees, a minimum leaf size of eight, and seed 42 were fitted for the clinical and eight pathway models and evaluated in every internal and external evaluation, with the clinical Cox model as reference. Mutation models were fitted in TCGA only and added mutation burden, recurrent gene indicators, or mutation features combined with pathways to the clinical model. Penalty selection by cross validated C index over a grid of 0.001 to 10, and models with fixed penalties on unscaled covariates, were compared with the primary approach in a supplementary analysis (Table S2). The primary criterion was changed to the partial log likelihood after C index based tuning selected the largest penalty of its grid in 48% to 84% of folds.

Internal evaluation used ten repeats of stratified fivefold cross validation, with seed 42. Discrimination was measured using Harrell's C index (Harrell et al., 1982), calculated from out of fold predictions for each repeat and averaged across repeats. Prediction error was assessed using the integrated Brier score (IBS, (Graf et al., 1999)) over 12 to 60 months. The evaluation grid was shortened when fewer than ten patients remained at risk. The

censoring distribution was estimated in the evaluated cohort, and the IBS was scaled against a Kaplan Meier reference. Confidence intervals for model differences were obtained using patient level bootstrap resampling, with 1,000 resamples for internal evaluation and 2,000 for external evaluation. All models were evaluated on the same resampled patients to preserve paired comparisons.

For external validation, models were fitted in the full training cohort and applied without refitting to the other cohort, in both directions (Royston & Altman, 2013). Performance was assessed using the C index, IBS based on raw predicted survival, calibration slope, and calibration at 36 months. Calibration slopes were estimated by fitting a Cox model with the frozen linear predictor in the test cohort. Calibration at 36 months compared mean predicted risk with Kaplan Meier risk across fifths of predicted risk. A separate diagnostic analysis recalibrated the frozen linear predictor in the test cohort before calculating the IBS. These results were reported separately because they use test cohort information. Incremental association was also assessed separately from frozen predictive performance. In the test cohort, a likelihood ratio test compared a Cox model containing the frozen clinical predictor with one that also included the frozen pathway component, calculated using the training coefficients. Both coefficients were estimated in the test cohort for this one degree of freedom test, which assessed additional association beyond the clinical predictor, not the performance of the fully frozen model.

The primary comparison family comprised internal validation in TCGA, internal validation in CGGA grades II and III, and external validation in both directions between TCGA and CGGA grades II and III. Each setting included the three pathway Cox models, giving 12 comparisons. Holm adjustment (Holm, 1979) was applied separately to the C index comparisons, the IBS comparisons, and the six association tests. Analyses including CGGA grade IV tumors were secondary and unadjusted. Restricting the reverse training analysis to CGGA grades II and III was decided after an initial run because TCGA contained no grade IV tumors. Both reverse analyses were retained, and this post hoc decision was documented.

Pooled association analyses combined TCGA and CGGA grades II and III. They were used to assess associations, not predictive performance. Five principal components of the 50 pathway scores were calculated without using outcome data. A Cox model adjusted for age, grade, and molecular subtype and stratified by cohort and CGGA batch was compared with the same model plus the five components using a likelihood ratio test with five degrees of freedom. Component by cohort interactions assessed heterogeneity. Analyses were repeated across all grades and in CGGA patients with available MGMT status. Additional covariates were mean log2 expression across genes and the number of detected genes, defined as genes with log2 TPM above 1. The proportional hazards assumption was assessed using a Schoenfeld type test based on rank transformed time. Nonlinearity in grade was tested using a likelihood ratio test.

To assess pathway specificity, each Hallmark set was compared with 1,000 random gene sets matched for size and mean expression profile. The pathway z statistic was estimated in a Cox model adjusted for age, grade, and subtype and stratified by cohort and batch. Empirical P values were calculated from the two sided random set distribution, with a minimum value of 0.002, and adjusted across 50 pathways using the Benjamini Hochberg procedure (Benjamini & Hochberg, 1995). A pathway was considered more prognostic than random sets in both cohorts if its empirical P value was below 0.05 in each cohort, without adjustment for multiplicity, and the association direction agreed. Adjusted hazard ratios per standard deviation were estimated separately by cohort

with Benjamini Hochberg adjustment across the 50 pathways, and agreement between cohorts was summarized using Spearman correlation.

A broader random set benchmark replaced each Hallmark set with a set of the same size sampled to match its distribution across 20 mean expression bins. The five component analysis and likelihood ratio test were repeated across 200 random gene set families. The observed statistic was compared with the empirical distribution of random family statistics. A second benchmark used the first five principal components of all shared genes. Another analysis stratified baseline hazards by cohort, batch, and molecular subtype and used 500 random families.

Finally, a pathway score was constructed in one cohort and tested in the other. The eight pathways with the largest excess z statistics over the random set mean, measured in units of the random set standard deviation, were selected in the training cohort. Their standardized scores were averaged to form the combined score. In the other cohort, a one degree of freedom likelihood ratio test assessed the score's association with survival after adjustment for age, grade, and subtype. A clinical model with the score was also fitted in the training cohort and applied unchanged to the test cohort. The change in C index was estimated with a bootstrap confidence interval and compared with 1,000 random score constructions using the same gene set sizes and expression profiles.

Analyses were conducted in Python 3.11.4 using scikit-learn 1.9.0, scikit-survival 0.28.0 (Pölsterl, 2020), lifelines 0.30.3 (Davidson-Pilon, 2019), GSEApy 1.3.1 (Fang et al., 2023), NumPy 1.26.4, pandas 3.0.5, SciPy 1.17.1 and Matplotlib 3.11.1. Random seeds were fixed at 42 unless otherwise specified. Batch was used for stratification in pooled association, pathway specificity, pathway hazard ratio, and pathway score analyses, but was not included as a predictor in the survival models. Tables and figures were generated programmatically from the analysis outputs.

Results

Cohorts and clinical model performance

The analytic cohorts included 352 TCGA patients with 87 deaths and 561 CGGA patients with 329 deaths (Table 1 and Figure S1). From the initial RNA sequencing cohorts, 13 of 365 TCGA patients and 69 of 630 CGGA patients were excluded. The main reason was unresolved molecular subtype (7 of 13 exclusions in TCGA and 64 of 69 in CGGA). Median follow up by the reverse Kaplan Meier method was 26.4 months in TCGA and 89.6 months in CGGA. Median age was 41.2 years in TCGA and 43.0 years in CGGA. The CGGA grade II and III subset included 354 patients and 152 deaths. Unlike TCGA, CGGA included 207 grade IV tumors. Molecular subtype distributions are reported in Table 1.

The clinical model achieved a cross validated C index of 0.852 in TCGA, 0.774 in CGGA grades II and III, and 0.789 in CGGA across all grades. Its IBS in TCGA was 0.114, compared with 0.178 for the Kaplan Meier reference. In external validation, the C index was 0.747 when the model was trained in TCGA and tested in CGGA grades II and III. In the reverse direction, it was 0.820. The corresponding calibration slopes were 0.64 and 1.16. Recalibration left the IBS unchanged at 0.129 in the forward direction and reduced it from 0.128 to 0.127 in the reverse direction. Including all CGGA grades gave C indices of 0.782 in the forward direction and 0.813 in the reverse direction (Figure 4).

Incremental prediction performance

Adding pathway scores produced small changes in discrimination across the four primary evaluation settings. Changes in C index ranged from −0.017 to +0.015 (Figure 1 and Table S1). No difference remained statistically significant after Holm adjustment across the 12 primary comparisons. The largest increase was observed for the eight pathway model trained in CGGA grades II and III and tested in TCGA, with a change of +0.015 (95% confidence interval 0.003 to 0.028, adjusted P = 0.084).

Changes in IBS were also generally small. Most comparisons showed little change in prediction error, although the eight pathway model in the reverse external validation reduced IBS by 0.0064 after adjustment for multiple testing (Holm P = 0.048). Secondary analyses including grade IV produced C index changes between −0.010 and +0.009.

Random survival forests did not consistently outperform the clinical Cox model. Across internal and external evaluations, changes in C index ranged from −0.018 to +0.021 for the clinical forests and from −0.069 to +0.044 for the eight pathway forests. The eight pathway forest exceeded the clinical Cox model in the reverse external evaluation, with increases of +0.005 for CGGA grades II and III and +0.044 for all CGGA grades. Mutation burden, recurrent gene mutations, and the combination of pathways and mutations did not improve discrimination in TCGA cross validation. Their C index changes were −0.004, −0.003, and −0.002, respectively (Table S3).

Within TCGA molecular subtypes, the clinical model had C indices of 0.915 in IDH mutant codeleted tumors, 0.643 in IDH mutant noncodeleted tumors, and 0.683 in IDH wild type tumors. No pathway model improved discrimination in any subtype. All within subtype confidence intervals for C index changes in the two primary external evaluations included zero.

Results were similar in sensitivity analyses. For progression free interval in TCGA, the clinical C index was 0.723, and all three pathway models had lower C indices, with changes ranging from −0.009 to −0.005. Using ssGSEA scores in TCGA also produced negative changes. Standardizing genes within the test cohort and sequencing batch yielded changes between −0.009 and +0.0003 in the forward direction and between +0.001 and +0.012 in the reverse direction (Table S3).

The eight pathway selection procedure most frequently selected interferon gamma response, IL 6/JAK/STAT3 signaling, interferon alpha response, inflammatory response, and IL 2/STAT5 signaling. The ridge penalty reached the upper limit of its original grid in 6% of folds for the eight pathway model, 34% for the 50 pathway model, and 66% for the five component model (Figure S2).

Extending the penalty grid to 3,000 changed each primary C index difference by less than 0.0002. When pathway selection was repeated within each inner cross validation split, the reverse external improvement for the eight pathway model was not reproduced. The corresponding C index changes were −0.002 in TCGA cross validation, −0.005 in CGGA cross validation, −0.001 in the forward external evaluation, and +0.009 in the reverse evaluation. The association of the frozen eight pathway component with survival remained similar in the forward direction (likelihood ratio chi square = 6.65, P = 0.0099) but was weaker in the reverse direction (chi square = 1.78, P = 0.18). In the reverse direction, the IBS of the nested eight pathway model (0.117) remained below that of the clinical model (0.128), although this difference was not tested.

Additional survival information from pathway scores

In external validation, some pathway components were associated with survival beyond the frozen clinical predictor when their coefficients were re estimated in the test cohort (Table 2). In CGGA grades II and III, the likelihood ratio tests gave chi square values of 6.76 for eight pathways (P = 0.0093, Holm adjusted P = 0.046), 4.74 for 50 pathways (P = 0.029, adjusted P = 0.077), and 11.98 for five components (P = 0.0005, adjusted P = 0.0032). In TCGA, the corresponding values were 6.80 (P = 0.0091, adjusted P = 0.046), 4.99 (P = 0.026, adjusted P = 0.077), and 1.72 (P = 0.19). The frozen five component predictor varied little in CGGA grades II and III, with a standard deviation of 0.0025.

In pooled analyses of 706 patients with 239 deaths, the five pathway components added information to age, grade, and molecular subtype (chi square = 25.08, 5 degrees of freedom, P = 0.00013). The component effects differed between cohorts (interaction chi square = 13.03, P = 0.023). The corresponding chi square values were 8.79 in TCGA (P = 0.12) and 25.88 in CGGA (P = 0.0001). Across all grades, the chi square was 31.68 (P = 0.000007). In CGGA patients with available MGMT status, the association remained after adjustment for MGMT promoter methylation (chi square = 20.34, P = 0.0011, n = 292).

The first five principal components explained 80.1%, 7.2%, 3.8%, 2.2%, and 1.2% of the variance in pathway scores (Figure S3). In pooled analyses, component 1 had a hazard ratio of 1.19 per standard deviation (95% confidence interval 1.06 to 1.35), and component 3 had a hazard ratio of 1.28 (1.12 to 1.47). Component 1 correlated strongly with mean gene expression in TCGA (Spearman correlation 0.94) and CGGA (0.67). Correlations with the number of detected genes were 0.62 and 0.58, respectively. Adding mean expression and detected gene count to the clinical model provided limited additional information (chi square = 5.51, 2 degrees of freedom, P = 0.064). The pathway component association remained after adjustment for these measures (chi square = 26.46, P = 0.00007).

Component 3 loaded positively on E2F targets, G2 M checkpoint, and Myc targets and negatively on KRAS signaling down and bile acid metabolism. Component 2 showed different associations between cohorts, with a hazard ratio of 0.77 in TCGA and 1.03 in CGGA. Component specific results are shown in Figure 2A.

Comparison with random gene sets

The Hallmark components did not outperform random gene set families in the pooled benchmark. Across 200 random families, the five component likelihood ratio statistic had a mean of 31.5, a median of 31.3, and a 95th percentile of 47.2, compared with 25.1 for Hallmark components. The empirical P value was 0.77. After adjustment for generic expression covariates, the median statistic for random families was 28.9, compared with 26.5 for Hallmark components (empirical P = 0.64).

The first five principal components of all 15,622 shared genes produced a chi square statistic of 40.1 ($P = 1.4 \times 10^{-7}$), increasing to 42.7 after adjustment for generic covariates. When the baseline hazard was stratified by cohort, sequencing batch, and molecular subtype, the Hallmark component statistic was 18.8 (P = 0.0021), compared with 25.7 for whole transcriptome components (P = 0.0001). The mean statistic across 500 random families was 18.7, and the empirical P value for a random family at least as informative as Hallmark was 0.45. The cohort specific analysis was not significant in TCGA for Hallmark components (chi square = 5.5, P = 0.36), whereas the association remained in CGGA (chi square = 20.2, P = 0.0011). Random gene set statistics were similar to the Hallmark statistic in both cohorts.

The proportional hazards assumption was not met for the IDH mutant indicators in CGGA grades II and III or TCGA. A nonlinear grade term did not improve the model in CGGA across all grades (P = 0.18).

Pathway level results

After adjustment for multiple testing, no individual pathway was associated with survival in TCGA. Two pathways reached a false discovery rate below 0.05 in CGGA grades II and III, and two did so in CGGA across all grades. None replicated across cohorts. Hazard ratios were directionally consistent for 46 of 50 pathways in CGGA grades II and III, but their overall correlation was modest (Spearman correlation 0.27, P = 0.061). Across all CGGA grades, 45 of 50 effects had the same direction (correlation 0.35, P = 0.012). No pathway reached the adjusted significance threshold within any molecular subtype.

Compared with 1,000 random gene sets matched for size and expression profile, 20 Hallmark pathways were more prognostic and two were less prognostic in the pooled analysis after false discovery rate adjustment (Figure 3). Sixteen pathways were more prognostic than random sets in both cohorts at nominal P below 0.05. These included E2F targets, G2 M checkpoint, interferon alpha response, interferon gamma response, epithelial mesenchymal transition, IL 6/JAK/STAT3 signaling, apoptosis, coagulation, TNF alpha signaling via NF kappa B, mitotic spindle, mTORC1 signaling, TGF beta signaling, complement, p53 pathway, hypoxia, and allograft rejection.

The highest pooled hazard ratios per standard deviation were observed for E2F targets (1.34), G2 M checkpoint (1.32), interferon alpha response (1.32), interferon gamma response (1.29), epithelial mesenchymal transition (1.26), and IL 6/JAK/STAT3 signaling (1.25). Oxidative phosphorylation and KRAS signaling down were less prognostic than matched random sets in both cohorts. Random gene sets also showed positive associations with mortality, with median mean z statistics of 2.3 in the pooled data, 0.5 in TCGA, and 1.2 in CGGA.

Cross cohort evaluation of selected pathway scores

The score selected in TCGA comprised interferon gamma response, TNF alpha signaling via NF kappa B, IL 2/STAT5 signaling, complement, apoptosis, epithelial mesenchymal transition, interferon alpha response, and mTORC1 signaling. In CGGA grades II and III, the score was associated with survival, with a hazard ratio of 1.19 per standard deviation (95% confidence interval 1.01 to 1.39, likelihood ratio chi square = 4.89, P = 0.027). However, the C index changed by −0.023 (95% confidence interval −0.045 to −0.001, bootstrap P = 0.038).

The score selected in CGGA comprised E2F targets, G2 M checkpoint, Myc targets V1, interferon gamma response, interferon alpha response, epithelial mesenchymal transition, TNF alpha signaling via NF kappa B, and mitotic spindle. In TCGA, it was associated with survival, with a hazard ratio of 1.32 (1.03 to 1.71, chi square = 4.62, P = 0.032). The change in C index was +0.014 (95% confidence interval −0.002 to 0.032, P = 0.079). In both directions, the observed score statistic exceeded those of all 1,000 random constructions. The empirical P value was 0.001 in each direction, the smallest value possible with 1,000 random constructions.

Penalty selection criteria

When the ridge penalty was selected by cross validated C index, it reached the limit of its grid in 24 to 42 of 50 folds for models containing pathway or mutation terms. The primary models selected the penalty by cross validated partial log likelihood over a wider grid.

Models with fixed penalties on unscaled covariates had a scaled Brier score of 0.19, compared with 0.36 for the unpenalized clinical model in TCGA cross validation. The fixed penalty clinical model had a calibration slope of 2.15 in the forward external evaluation (Table S2).

Discussion

Hallmark pathway scores did not consistently improve survival prediction beyond age, WHO grade, and molecular subtype. Changes in discrimination were small, varied in direction across cohorts, and did not remain significant after adjustment for the primary comparisons. The only adjusted improvement in prediction error occurred in the reverse external validation. The increase in C index for the eight pathway model was not reproduced when pathway selection was repeated within inner cross validation, although the lower prediction error in that direction persisted (0.117 versus 0.128, not tested). These findings indicate that pathway scores can carry information about survival without providing a consistent improvement in individual risk prediction.

The results also help place molecular prognostic models in context. Du et al. reported AUC values of 0.817 in training and 0.819 in testing for a multiomics model of lower grade glioma (Du et al., 2024). However, the performance of a complete model does not establish how much information its molecular features add beyond established clinical predictors. Our clinical model achieved a C index of 0.852 in TCGA, highlighting the importance of using a comparable clinical reference when evaluating new molecular predictors. Direct numerical comparison is inappropriate because the studies used different prediction settings and performance measures. Age, grade, and molecular subtype already capture important prognostic differences among adult diffuse gliomas, consistent with the current molecular classification (Eckel-Passow et al., 2015; Louis et al., 2021). Pathway scores may consequently contribute information that overlaps with these established predictors.

The pathway level analyses identified associations that were stronger than those of matched random sets. Sixteen pathways were more prognostic than matched random gene sets in both cohorts at nominal significance, including pathways related to proliferation, interferon signaling, inflammation, apoptosis, and epithelial mesenchymal transition. Scores constructed from selected pathways in one cohort were also associated with survival in the other. However, these associations did not translate into consistent gains in discrimination. One possible explanation is that pathway activity reflects biological differences already partly captured by grade and molecular subtype. The results support further investigation of selected pathways, but they do not establish that these pathways provide clinically useful prognostic improvement.

The random gene set benchmarks were important for interpreting these findings. Venet et al. showed that more than 90% of random signatures of over 100 genes were significantly associated with breast cancer outcome, and that 28 of 47 published signatures (60%) were not significantly better predictors than random signatures of identical size (Venet et al., 2011). Our analyses extend this concern to lower grade glioma. Random gene sets showed substantial associations with mortality, while whole transcriptome components provided at least as much information as the Hallmark components in the pooled analyses. The first pathway component was also strongly correlated with overall expression level and the number of detected genes. These patterns suggest that some prognostic information reflects broad expression characteristics rather than pathway specific biology. This does not mean that the Hallmark pathways lack biological relevance. The random sets were matched for size and mean expression, but not for coexpression structure. The benchmark therefore tests whether curated pathways

outperform a defined set of controls, not whether they have no biological function. This distinction is consistent with the methodological concerns about the different questions addressed by gene set tests (Goeman & Bühlmann, 2007).

The external validation analyses also showed why discrimination and calibration need to be evaluated separately. A model trained in TCGA had a calibration slope of 0.64 when applied to CGGA grades II and III, indicating that its predictions were too extreme. The reverse model had a slope of 1.16, closer to the ideal value of one. Differences in follow up, patient composition, sequencing platforms, and batch structure may have contributed to these patterns. Recalibration improved prediction error only slightly and used information from the test cohort, so it does not represent the performance of a completely frozen model. Similarly, the likelihood ratio tests showed that some transferred pathway components remained associated with survival when their coefficients were re estimated in the test cohort. These tests assess additional association beyond the clinical predictor, not the predictive performance of the fully frozen model. This distinction is important when interpreting external validation results (Ramspek et al., 2021; Royston & Altman, 2013).

Several limitations should be considered. TCGA included 87 deaths and had a median follow up of 26.4 months, limiting the precision of some estimates. Treatment information was unavailable, and MGMT promoter methylation was available only in CGGA. WHO grade was modeled as an ordinal term, and the proportional hazards assumption was violated for the IDH mutant indicators. The cohorts also differed in grade distribution, follow up, and sequencing characteristics. Although the stratified analyses supported the main findings, these differences limit direct comparison between cohorts. The random gene set benchmarks did not preserve coexpression structure, and only the Enrichr 2020 Hallmark collection was evaluated. In addition, the eight pathways were selected before penalty tuning, the penalty criterion was changed from the C index to the partial log likelihood after a first run, and the restriction of the reverse training analysis to CGGA grades II and III was decided after that run. Sensitivity analyses addressed pathway selection and penalty tuning, but these design choices should be considered when interpreting the results. Mutation models could not be externally validated because they were fitted in TCGA only.

The main strengths were the use of two independent public cohorts, validation in both directions, assessment of calibration and censoring adjusted prediction error, a primary comparison family fixed before the results were known, and direct benchmarking against random gene sets and whole transcriptome components. Together, these analyses show that pathway associations alone are insufficient to establish incremental prognostic value. Future studies of molecular survival models should compare the added features with a well specified clinical model, use independent validation with frozen predictions, and evaluate whether curated signatures outperform appropriate random controls. Such evaluation is needed before pathway based signatures can be considered useful additions to clinical prognostic assessment.

Declarations

Supplementary material

The supplementary file contains Figure S1 (cohort flow), Figure S2 (pathway selection stability), Figure S3 (pathway component loadings), Table S1 (primary model comparisons), Table S2 (penalty selection criteria), and Table S3 (sensitivity analyses).

Acknowledgments

We thank the teams responsible for the Cancer Genome Atlas, the Genomic Data Commons, the Chinese Glioma Genome Atlas, and cBioPortal for providing access to the data used in this study. We also acknowledge AREF for its support.

Ethics approval

This study used publicly available, deidentified data from TCGA and CGGA. No new data were collected directly from human participants. Ethics approval and informed consent were not required for this secondary analysis of deidentified data, subject to the applicable institutional requirements.

Data availability

The original TCGA and CGGA datasets are publicly available through their respective repositories, and molecular subtype data are available through cBioPortal. Processed data and analysis outputs are available from the corresponding authors upon reasonable request.

Author contributions

A. D., A. K.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing. M.S.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft. All authors read and approved the final manuscript.

Funding

This study received financial support from Pathfinder Epidemiology Academy.

References

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F.,…Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, *462*(7269), 108-112. https://doi.org/10.1038/nature08460

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brat, D. J., Verhaak, R. G., Aldape, K. D., Yung, W. K., Salama, S. R., Cooper, L. A.,…Radenbaugh, A. J. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med*, *372*(26), 2481-2498. https://doi.org/10.1056/NEJMoa1402121

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B.,…Logullo, P. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *Bmj*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Davidson-Pilon, C. (2019). lifelines: survival analysis in Python. *Journal of Open Source Software*, *4*, 1317. https://doi.org/10.21105/joss.01317

Du, Z., Jiang, Y., Yang, Y., Kang, X., Yan, J., Liu, B., & Yang, M. (2024). A multi-omics analysis-based model to predict the prognosis of low-grade gliomas. *Scientific Reports*, *14*(1), 9427. https://doi.org/10.1038/s41598-024-58434-8

Eckel-Passow, J. E., Lachance, D. H., Molinaro, A. M., Walsh, K. M., Decker, P. A., Sicotte, H.,…Jenkins, R. B. (2015). Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. *N Engl J Med*, *372*(26), 2499-2508. https://doi.org/10.1056/NEJMoa1407279

Fang, Z., Liu, X., & Peltz, G. (2023). GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics*, *39*(1), btac757. https://doi.org/10.1093/bioinformatics/btac757

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, *23*(8), 980-987. https://doi.org/10.1093/bioinformatics/btm051

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*, *18*(17-18), 2529-2545. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5

Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543-2546.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841-860. https://doi.org/10.1214/08-aoas169

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, *1*(6), 417-425. https://doi.org/10.1016/j.cels.2015.12.004

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D.,…Hu, H. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, *173*(2), 400-416.e411. https://doi.org/10.1016/j.cell.2018.02.052

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D.,…Ellison, D. W. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol*, *23*(8), 1231-1251. https://doi.org/10.1093/neuonc/noab106

Pölsterl, S. (2020). scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, *21*, 1-6.

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2021). External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*, *14*(1), 49-58. https://doi.org/10.1093/ckj/sfaa188

Royston, P., & Altman, D. G. (2013). External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*, *13*, 33. https://doi.org/10.1186/1471-2288-13-33

Venet, D., Dumont, J. E., & Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*, *7*(10), e1002240. https://doi.org/10.1371/journal.pcbi.1002240

Zhao, Z., Zhang, K. N., Wang, Q., Li, G., Zeng, F., Zhang, Y.,…Jiang, T. (2021). Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients. *Genomics Proteomics Bioinformatics*, *19*(1), 1-12. https://doi.org/10.1016/j.gpb.2020.10.005

Figures

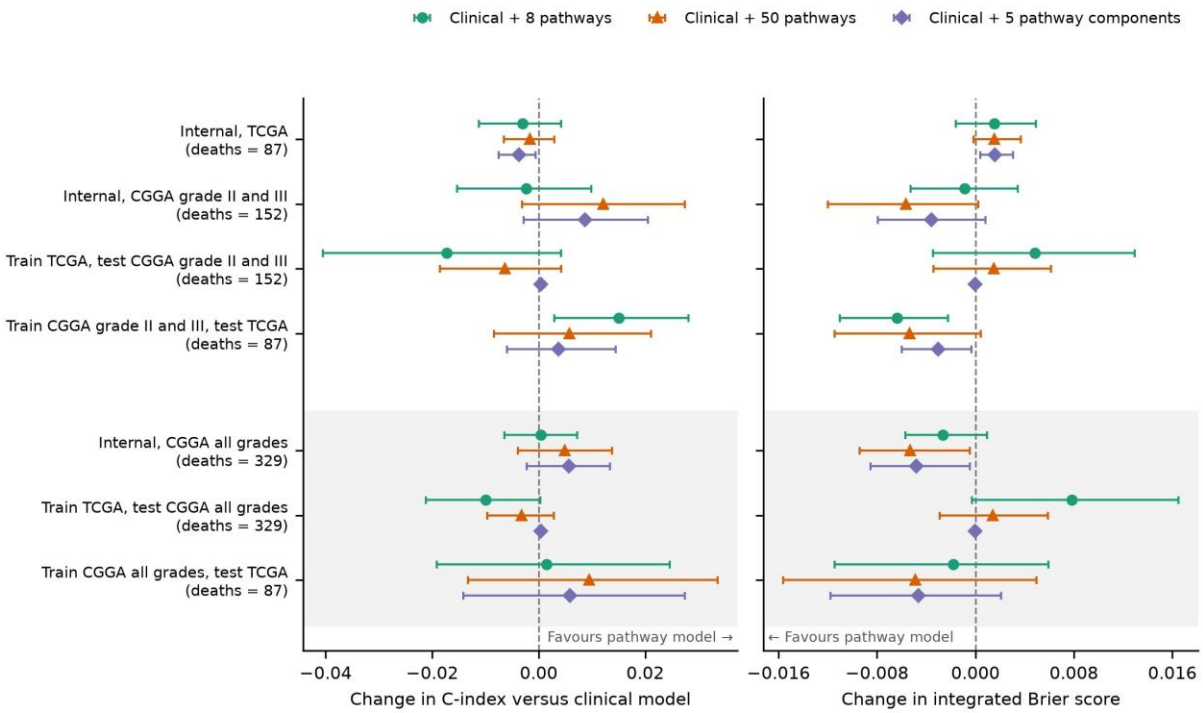


Figure 1. Change in C-index (left) and integrated Brier score (right) when pathway scores are added to the clinical model. Four primary evaluations are shown above the shaded area and three secondary evaluations that include grade IV within it. Bars are 95% patient level bootstrap confidence intervals.

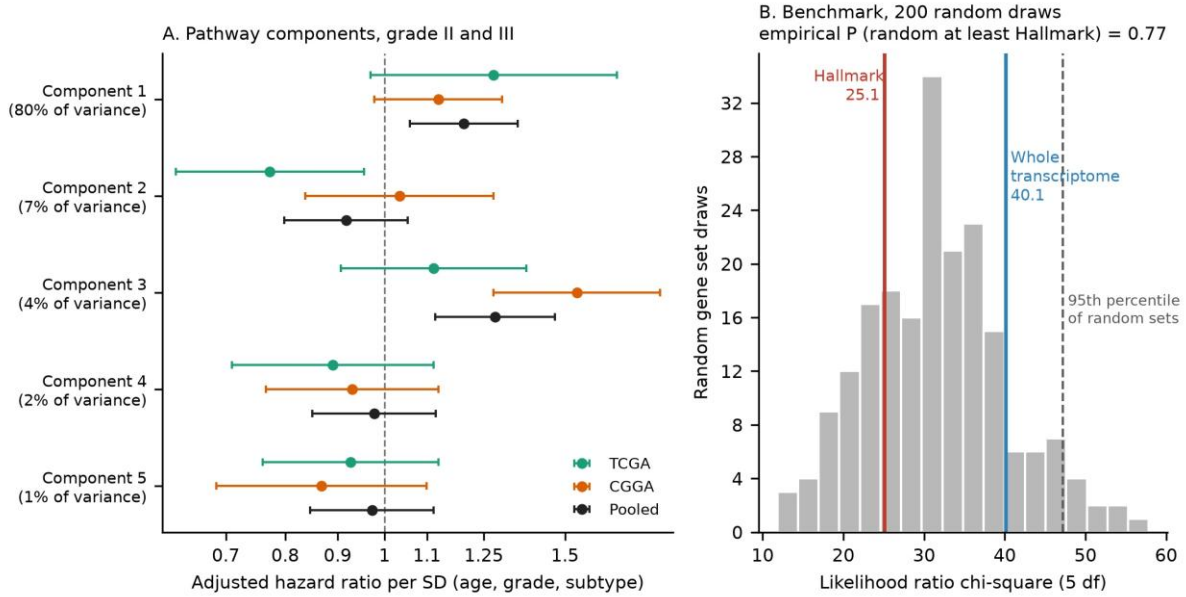


Figure 2. Pathway components and random gene set benchmark in TCGA and CGGA grade II and III. A, adjusted hazard ratios per SD of the five pathway components. B, likelihood ratio chi-square (5 df) of 200 random gene set families, with the values for the Hallmark sets and for whole transcriptome components.

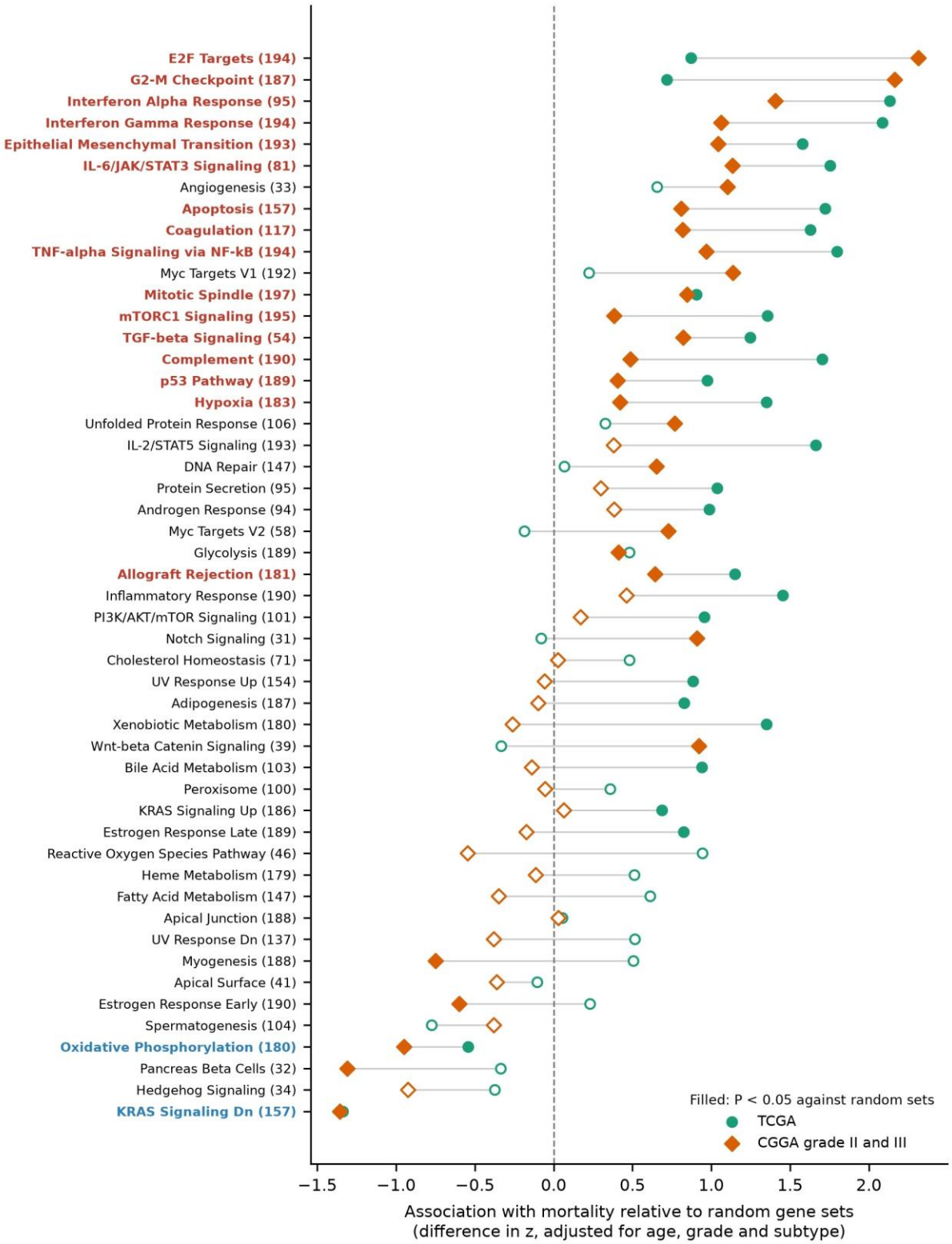
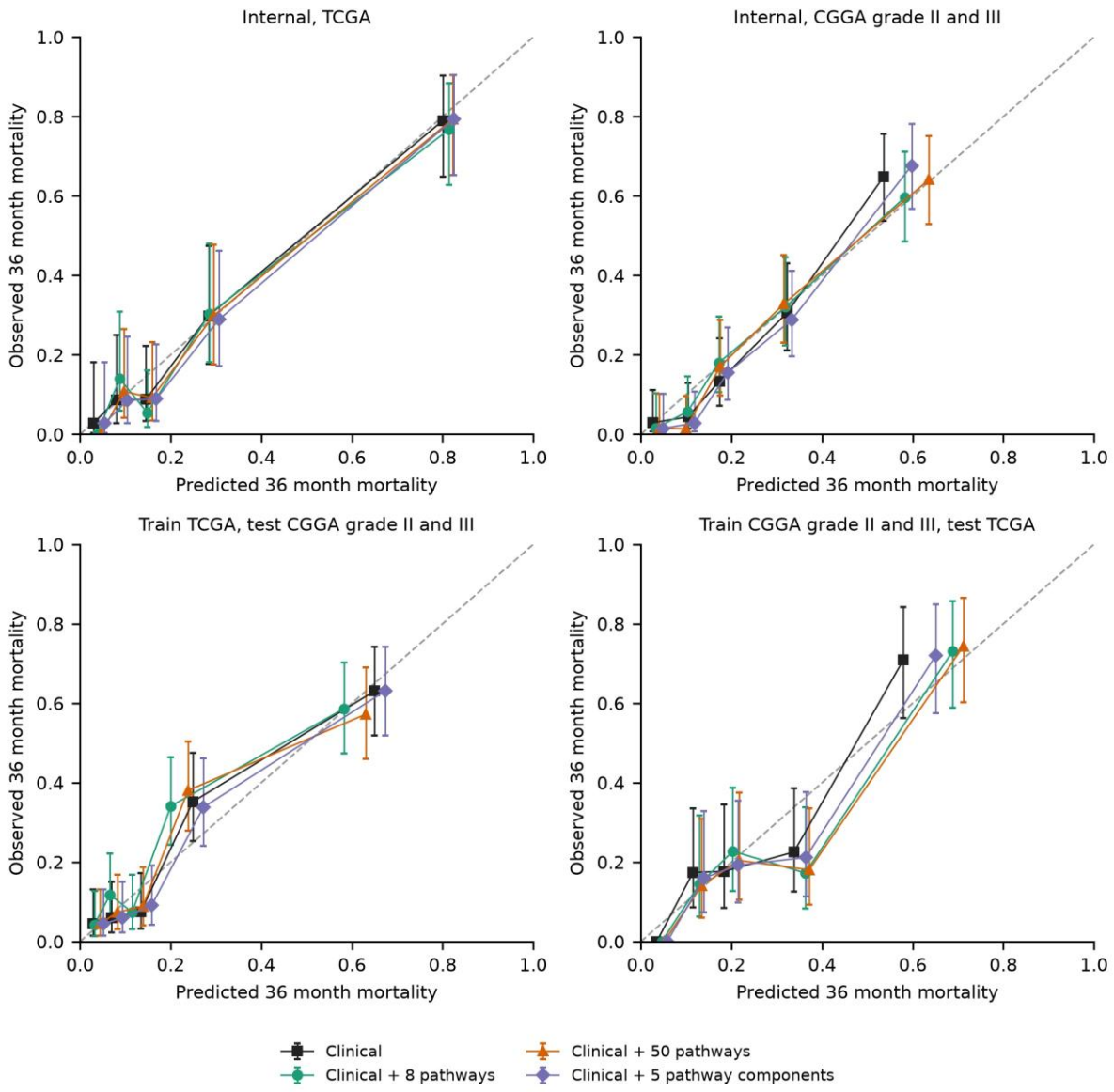


Figure 3. Association with mortality of 50 Hallmark pathways relative to random expression matched gene sets of the same size (difference in z, adjusted for age, grade and subtype). Filled symbols, P below 0.05 against 1000 random sets. Bold labels, P below 0.05 against random sets in both cohorts with the same direction, without adjustment for the two cohorts (red, more prognostic than random sets, blue, less).



469

470 **Figure 4.** Predicted and observed 36 month mortality by risk group for the clinical model and three pathway
471 models in four primary evaluations. Bars are 95% confidence intervals.

472

Tables

Table 1. Baseline characteristics of the TCGA and CGGA analytic cohorts.

Characteristic	TCGA-LGG (n = 352)	CGGA, WHO grade II and III (n = 354)	CGGA, all grades (n = 561)
Age at diagnosis, years, median (IQR)	41.2 (33.3, 54.4)	40.0 (34.0, 46.0)	43.0 (36.0, 54.0)
Female sex, n (%)	155 (44.0%)	150 (42.4%)	231 (41.2%)
WHO grade, n (%)			
Grade II	160 (45.5%)	196 (55.4%)	196 (34.9%)
Grade III	192 (54.5%)	158 (44.6%)	158 (28.2%)
Grade IV	0	0	207 (36.9%)
Molecular subtype, n (%)			
IDH-wildtype	66 (18.8%)	99 (28.0%)	278 (49.6%)
IDH-mutant, non-codeleted	174 (49.4%)	142 (40.1%)	168 (29.9%)
IDH-mutant, 1p/19q-codeleted	112 (31.8%)	113 (31.9%)	115 (20.5%)
Outcome			
Deaths, n (%)	87 (24.7%)	152 (42.9%)	329 (58.6%)
Median follow-up, months (reverse Kaplan-Meier)	26.4	91.3	89.6

IQR, interquartile range. Follow-up was estimated with the reverse Kaplan-Meier method. CGGA, Chinese Glioma Genome Atlas. TCGA-LGG, The Cancer Genome Atlas lower-grade glioma cohort.

Table 2. Association of pathway scores with survival beyond clinical terms. Panel A, frozen pathway component on the frozen clinical predictor in the test cohort. Panel B, pooled unsupervised pathway components.

A. Frozen pathway component added to the frozen clinical predictor in the test cohort

Model	n	Deaths	Chi-square	df	P	Holm P
Train TCGA, test CGGA grade II and III						
Clinical + 8 pathways	354	152	6.76	1	0.009	0.046
Clinical + 50 pathways	354	152	4.74	1	0.029	0.077
Clinical + 5 pathway components	354	152	11.98	1	<0.001	0.003
Train CGGA grade II and III, test TCGA						
Clinical + 8 pathways	352	87	6.80	1	0.009	0.046
Clinical + 50 pathways	352	87	4.99	1	0.026	0.077
Clinical + 5 pathway components	352	87	1.72	1	0.189	0.189

Likelihood ratio test in the test cohort. The coefficients of the frozen clinical predictor and of the frozen pathway component were both estimated in the test cohort, so this is a test of association and not a measure of the performance of the completely frozen model. Holm adjustment covers the six tests. df, degrees of freedom.

B. Pooled unsupervised pathway components

Test	n	Deaths	Chi-square	df	P
TCGA + CGGA WHO II/III					
5 PCs added to age, grade, subtype (strata cohort and batch)	706	239	25.08	5	<0.001
cohort heterogeneity (PC x cohort interactions)	706	239	13.03	5	0.023
TCGA only, pooled PCs	352	87	8.79	5	0.118
CGGA only, pooled PCs	354	152	25.88	5	<0.001
TCGA + CGGA all grades					
5 PCs added to age, grade, subtype (strata cohort and batch)	913	416	31.68	5	<0.001
cohort heterogeneity (PC x cohort interactions)	913	416	20.23	5	0.001
TCGA only, pooled PCs	352	87	8.64	5	0.124
CGGA only, pooled PCs	561	329	37.49	5	<0.001
CGGA WHO II/III (patients with MGMT known)					
CGGA-only PCs, without MGMT	292	130	18.94	5	0.002
CGGA-only PCs, with MGMT	292	130	20.34	5	0.001
CGGA all grades (patients with MGMT known)					
CGGA-only PCs, without MGMT	483	294	33.07	5	<0.001
CGGA-only PCs, with MGMT	483	294	34.52	5	<0.001

Cox models stratified by cohort and batch with age, WHO grade and molecular subtype. Components are the first five principal components of the 50 pathway scores, computed without reference to outcome.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx