APOLLO: An accurate and independently validated prediction model of lower-grade gliomas overall survival and a comparative study of model performance

Jiajin Chen, Sipeng Shen, Y Li, Juanjuan Fan, Shiyou Xiong, Jingtong Xu, Chenxu Zhu, Lijuan Lin, Xuesi Dong, Weiwei Duan, Yang Zhao, Xu Qian, Zhonghua Liu, Yongyue Wei, David C. Christiani, Ruyang Zhang, Feng Chen

Summary

Background Virtually few accurate and robust prediction models of lower-grade gliomas (LGG) survival exist that may aid physicians in making clinical decisions. We aimed to develop a prognostic prediction model of LGG by incorporating demographic, clinical and transcriptional biomarkers with either main effects or gene-gene interactions.

Methods Based on gene expression profiles of 1,420 LGG patients from six independent cohorts comprising both European and Asian populations, we proposed a 3-D analysis strategy to develop and validate an Accurate Prediction Model of Lower-grade gliomas Overall survival (APOLLO). We further conducted decision curve analysis to assess the net benefit (NB) of identifying true positives and the net reduction (NR) of unnecessary interventions. Finally, we compared the performance of APOLLO and the existing prediction models by the first systematic review.

Findings APOLLO possessed an excellent discriminative ability to identify patients at high mortality risk. Compared to those with less than the 20th percentile of APOLLO risk score, patients with more than the 90th percentile of APOLLO risk score had significantly worse overall survival \( (HR=54.18, 95\% CI: 34.73-84.52, P=2.66 \times 10^{-6}) \).

*Corresponding author: Ruyang Zhang, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China.
**Corresponding author: Feng Chen, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China.
***Corresponding author: David C. Christiani Harvard School of Public Health and Harvard Medical School, Building I Room 1401, 665 Huntington Avenue, Boston, MA 02115, USA.
E-mail addresses: dchris@hsph.harvard.edu (D.C. Christiani), zhangruyang@njmu.edu.cn (R. Zhang), fengchen@njmu.edu.cn (F. Chen).
† These authors contributed equally to this work.
* Senior author who supervised the work.

eBioMedicine 2022;79: 104007
Published online xxx
https://doi.org/10.1016/j.ebiom.2022.104007
Further, APOLLO can accurately predict both 36- and 60-month survival in six independent cohorts with a pooled AUC_{36-month}=0.901 (95% CI: 0.879-0.923), AUC_{60-month}=0.843 (95% CI: 0.815-0.871) and C-index=0.818 (95% CI: 0.800-0.835). Moreover, APOLLO offered an effective screening strategy for detecting LGG patients susceptible to death (NB_{36-month}=0.166, NR_{36-month}=40.1% and NB_{60-month}=0.258, NR_{60-month}=19.2%). The systematic comparisons revealed APOLLO outperformed the existing models in accuracy and robustness.

Interpretation APOLLO has the demonstrated feasibility and utility of predicting LGG survival (http://bigdata.njmu.edu.cn/APOLLO).

Funding National Key Research and Development Program of China (2016YFE0204900); Natural Science Foundation of Jiangsu Province (BK20191354); National Natural Science Foundation of China (81973142 and 82103946); China Postdoctoral Science Foundation (2020M681671); National Institutes of Health (CA209414, CA249096, CA092824 and ES000002).

Copyright © 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Keywords: Lower-grade gliomas; Survival; Prognostic prediction; Nomogram; Online tool; Systematic review

Research in Context

Evidence before this study

We searched PubMed, Embase, MEDLINE, Web of Science, and Cochrane Library for articles about prognostic prediction models of LGG published before Aug 30, 2021, using search term "((lower-grade glioma) OR (lgg)) AND ((progn*) OR (survival)) AND ((predict*) OR (auc) OR (area under the curve) OR (receiver operator characteristic curve) OR (c-index) OR (c statistic) OR (roc) OR (calibration))". We found the existing models underwent limited prediction accuracy and model validation, as most of them either solely relied on training populations or retrained models in testing populations to assess the model performance, which might be overestimated due to overfitting. Additionally, most of these existing models have limited model robustness and transportability to accommodate independent populations, impeding their wide applications.

Added value of this study

In this study, we collected 1,420 LGG patients from six European and Asian populations and proposed an effective modeling strategy to develop and validate an Accurate Prediction mOdel of Lower-grade gLiomAs Overall survival (APOLLO), which has the demonstrated feasibility and utility in distinguishing LGG patients at high risk of mortality and predicting their survival. Our systematic review revealed that APOLLO outperformed the existing models in accuracy and robustness.

Implications of all the available evidence

APOLLO has clinical benefits at identifying LGG patients at high mortality risk and presents a higher net benefit of identifying true positives and net reduction of unnecessary interventions. A convenient online tool to implement APOLLO was developed at http://bigdata.njmu.edu.cn/APOLLO.

Introduction

Gliomas, the most common malignant cancer in the brain and central nervous system, account for over 80% of malignant brain tumors. Lower-grade gliomas (LGG), consisting of diffuse low- and intermediate-grade gliomas, are graded II and III by World Health Organization (WHO). Compared to those diagnosed as glioblastoma (GBM) with a WHO grade IV, LGG patients tend to have more favorable prognosis; however, 70% of them will progress to GBM within ten years. Thus, delaying tumor progression for LGG patients is critical. What is often overlooked is wide heterogeneity of LGG prognosis that is ubiquitous for those even with similar clinical features, indicating possible molecular underpinnings of the disease progression process. As a crucial milestone, the WHO Classification of Tumors of the Central Nervous System synthesized molecular and histological information to reclassify gliomas, by using well recognized molecular biomarkers. Recent evidence has emerged that gene expressions may pose inducible and reversible effects on LGG prognosis via several channels, including immunity, stemness, and autophagy. The prognostic prediction utilizing biomarkers can aid physicians in making clinical decisions or guiding adjuvant therapy. Recently, much effort has been shifting to the LGG prognostic prediction. However, existing prediction models have various technical bottlenecks, impeding their wide applications. Specifically, these models
underwent limited model validation, as most of them either solely relied on training populations or retrained models in testing populations to assess the model performance, which might be overestimated due to overfitting.\textsuperscript{9,14} Therefore, most of these existing models have limited model robustness and transportability to accommodate independent populations.\textsuperscript{15,16}

Furthermore, almost all of the studies merely focused on predictors with main effects, but neglected predictors exhibiting gene-gene (\(G \times G\)) interactions, which may provide pivotal clues regarding the biologic mechanisms of complex diseases\textsuperscript{17} and enhance prediction accuracy,\textsuperscript{18,19} as evidenced by our own study of lung cancer.\textsuperscript{20}

To address challenges in LGG survival prediction, we developed an Accurate and independently validated Prediction mOdel of Lower-grade gliomas Overall survival (APOLLO) which identifies and includes biomarkers with significant main effects or \(G \times G\) interactions, based on six cohorts with both European and Asian populations. Additionally, we have developed a free online tool implementing APOLLO to facilitate prediction of LGG survival.

Materials and Methods

Data collection and study population

We curated the clinical and gene expression data of LGG patients from six glioma cohorts, namely, the Cancer Genome Atlas (TCGA),\textsuperscript{21} the Chinese Glioma Genome Atlas (CGGA),\textsuperscript{22} CGGA2,\textsuperscript{23} Rembrandt (GSE168476),\textsuperscript{24} Weller (GSE61774),\textsuperscript{25} and Gravendeel (GSE16011) cohorts.\textsuperscript{26} Only newly diagnosed LGG patients with complete overall survival time and transcriptomics data were retained. With the focus on biological functions and clinical utility, we considered a total of 723 pan-cancer driving genes defined by COSMIC,\textsuperscript{27} among them, included in our study were 680 genes shared by all six cohorts. All gene expression levels were log\(_2\)-transformed and standardized before being passed into association analyses; see the Supplementary Methods for the details of sample quality control. Included in our subsequent analyses were a total of 1,420 LGG patients with 680 genes, whose demographic and clinical characteristics were summarized in Supplementary Table S1.

APOLLO construction and validation

Figure 1, depicting the study design and workflow, features a 3-D strategy (Double Types of Effects, Double Steps of Screening, and Double Steps of Modeling) for the development and validation of the APOLLO model.

(i) Double Types of Effects. For selection of important main effects and \(G \times G\) interactions, we considered Cox Models 1 and 2, respectively:

\[
\text{Model 1:} \quad h(t) = h_0(t) \exp(\alpha \times \text{gene} \times \sum \beta_i \times \text{covariate})
\]

\[
\text{Model 2:} \quad h(t) = h_0(t) \exp(\alpha \times \text{gene} \times \text{gene} \times \sum \beta_i \times \text{covariate})
\]

which adjusted for covariates, including age, WHO grade, \(IDH\) mutation and 1p/19q status (Supplementary Table S2).

(ii) Double Steps of Screening. We scanned the pan-cancer related genes to select candidate genes and interactions, and then validated them with an independent validation dataset. Specifically, on the TCGA cohort, we fitted Models 1 and 2 on each gene and interaction, respectively, and selected important genes and interactions by controlling the false positive rate at a 5% level (\(q\)-FDR≤5%). On the CGGA cohorts, we validated these selected genes or interactions: only those with \(P_{\text{corr}}<0.05\) and with same effect directions as in the discovery step would be selected as candidate biomarkers to be passed onto the next modeling stage.

(iii) Double Steps of Modeling. On the TCGA cohort and with the candidate genes and interactions identified from the previous screening stage, we used Cox models (adjusted for demographic and clinical predictors) to conduct forward stepwise regression, that is, using the likelihood ratio test with \(P_{\text{entry}}<0.05\) and \(P_{\text{remove}}>0.05\), to identify a final multivariable Cox model and construct APOLLO. As validation, we assessed the discriminative performance of the obtained APOLLO via area under the receiver operating characteristic curves (AUC) or concordance index (C-index) on one internal cohort (CGGA1) and four external cohorts, namely, CGGA2, Rembrandt, Weller and Gravendeel.

Bioinformatics analysis for transcriptional predictors

To understand the potential gene functions of the identified transcriptional predictors, we conducted a gene enrichment pathway analysis based on Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database by using R package \texttt{clusterProfiler}. Estimation of Stromal and Immune Cells in Malignant Tumor Tissues Using Expression Data (ESTIMATE)\textsuperscript{28} was used to predict the presence of stromal and immune cells in tumor tissue, and CIBERSORT was performed to determine the proportions of 22 immune cells from bulk tumors based on gene expression.\textsuperscript{29} Finally, the gene network analysis of screened genes and immune checkpoint genes was performed using GeneMANIA,\textsuperscript{30} a plugin of the Cytoscape application.

A systematic review of LGG survival prediction models

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Supplementary...
Table S3), we conducted a systematic literature search on prognostic prediction models of LGG using five major databases, namely, PubMed, Embase, MEDLINE, Web of Science, and Cochrane Library. The literature search and data extraction were done independently by two researchers (S.X. and J.X.), and the discrepancies were arbitrated by a third researcher (J.C.). Details of search strings, exclusion criteria and data extraction process were provided in Supplementary Methods. We totally retrieved 3,035 articles. After removing duplicates, 1,444 articles were included for further screening. Among them, 126 articles that met the criteria on title...
or abstract, were eligible for a full-text review. Finally, 54 articles fully meeting our selection criteria were retained and used for data extraction.

**APOLLO visualization and online software**

We generated a nomogram for visualizing APOLLO by using R package `rms`, which can be accessed at [http://bigdata.njmu.edu.cn/APOLLO](http://bigdata.njmu.edu.cn/APOLLO). With input values of predictors for a LGG patient, the online calculator immediately returns predicted survival rates and 95% confidence intervals (CIs) at any time point between 0 and 120 months, based on an interactive web-based Kaplan-Meier survival curve.

**Statistical analysis**

Continuous variables were summarized as mean ± standard deviation, and categorized variables were described by frequency (n) and proportion (%). Associations between characteristics and overall survival were evaluated by Cox models using R package `survival`. Study centers were adjusted for when analyzing the combined samples. Kaplan-Meier survival curves illustrated the survival differences across different risk groups. The prediction accuracy was presented by using a time-dependent receiver operating characteristic (ROC) curve and was assessed by the time-dependent AUC, which can be obtained from R package `timeROC`. We used calibration plots to evaluate the consistency between nomogram-predicted and observed risks, and conducted decision curve analysis (with details given in Supplementary Methods) to gauge the net benefit (NB) of identifying true high risk patients that ought to have intervention and the net reduction (NR) of unnecessary interventions, due to the use of APOLLO as a screening tool. Since these transcriptional predictors were validated by trans-ethnic populations, we assumed APOLLO had uniform and homogenous performance across cohorts. Thus, we performed meta-analysis to pool prediction accuracy of APOLLO from six cohorts using the fixed-effect model, implemented by the R package `meta`. Stratified analyses were displayed by forest plots using the R package `forestplot`.

APOLLO_Score = 0.0312 × age + 0.5276 × grade – 0.5510 × IDH – 0.5163 × 1p/19q + 0.7528 × Transcriptional_Score

Transcriptional_Score = 0.2976 × CHIC2 + 0.3500 × IGF2BP2 + 0.2387 × ITGAV + 0.5532 × MSN + 0.4034 × PLCG1 + 0.2361 × BCORL1 – 0.1082 × PRF1 – 0.2498 × BCORL1 × PRF1 – 0.1674 × HMGA1 – 0.1058 × TFG + 0.1930 × HMGA1 × TFG – 0.1922 × CTNND2 – 0.1814 × GOLGA5 – 0.2340 × CTNND2 × GOLGA5 – 0.0888 × FAS – 0.2073 × SMAD4 – 0.1724 × FAS × SMAD4

Statistical analyses were performed using R (version 3.6.3). A two-sided P value less than or equal to 0.05 was considered statistically significant unless otherwise specified. The source code and data were deposited at [https://github.com/JiajinChen/APOLLO](https://github.com/JiajinChen/APOLLO).

**Ethics**

The study was performed in accordance with Good Clinical Practice guidelines and the World Medical Association Declaration of Helsinki. All patients provided informed consents. All data used in this study were de-identified and no protected health data was needed.

**Role of the funding source**

The sponsors had no role in the study design, data collection, data analyses, interpretation, or writing of the study.

**Results**

**Development and construction of APOLLO**

First, 42 genes with main effects and 307 pairs of genes with G × G interactions were identified (q-FDR < 0.05) to be possibly associated with overall survival in TCGA cohort. Of them, 28 genes with main effects and 27 pairs of genes with G × G interactions were validated in CGGA1 cohort to be candidate transcriptional predictors (Supplementary Tables S4-S5). Then, out of these candidate transcriptional predictors and on the TCGA training cohort, we used forward stepwise regression strategy to construct a final Cox model, which included 5 genes with main effects and 4 pairs of genes with G × G interactions (Supplementary Table S6). Using the coefficient estimates from this final Cox model, APOLLO, which integrated demographic, clinical and transcriptional predictors, was defined as:

Transcriptional predictors of APOLLO and their immune relevance

KEGG enrichment analysis categorized gene probes into 30 pathways, including the glioma pathway, and GO annotation identified 279 biological process...
pathways, 24 molecular function pathways and 20 cellular component pathways, suggesting potential biological functions (Supplementary Table S7). We compared the proportions of 22 types of immune cells between high- and low-risk groups defined by the median transcriptional score (10.26889), and found that they were significantly different between the two groups (Supplementary Figure S1a). Further, the transcriptional score was correlated with the stromal, immune and ESTIMATE scores (Supplementary Figure S1b). Additionally, we observed high connectivity and large correlations between transcriptional predictors and immune checkpoint genes (Supplementary Figure S1c and Supplementary Figure S2a), indicating that the transcriptional predictors may play a role in immune responses. Numerous immunity-related drugs targeting these transcriptional predictors have been documented in the DrugBank database (Supplementary Table S8), and, thereby, APOLLO may have potential roles in guiding immunotherapy.

**Discriminative ability of APOLLO**

Patients in each of the six cohorts were categorized into low- and high-risk groups using the median APOLLO score (0.6943) obtained from the TCGA training set. The APOLLO score had an adequate discriminative ability in both training and testing sets. Compared to the low-risk group in the corresponding cohort, the high-risk group was associated with worse survival in the TCGA cohort (the training set) and CGGA1 (the internal testing cohort), exhibiting a large hazard ratio (HR) (HR_{CGGA1}=8.31, 95% CI: 3.47-14.8, \textit{P}=2.14 \times 10^{-16}; HR_{CGGA2}=4.86, 95% CI: 1.24-7.28, \textit{P}=1.78 \times 10^{-14}) (Figure 2a-b), and in the 4 external testing sets (HR_{CGGA2}=6.26, 95% CI: 2.86-13.68, \textit{P}=4.14 \times 10^{-6}; HR_{Rembrandt}=3.49, 95% CI: 2.06-5.91, \textit{P}=3.12 \times 10^{-4}; HR_{Weller}=3.41, 95% CI: 1.73-6.72, \textit{P}=3.90 \times 10^{-4}; HR_{Gravendeel}=2.19, 95% CI: 1.31-3.68, \textit{P}=2.88 \times 10^{-3}) (Figure 2c-d). We further illustrated the discriminative ability of the APOLLO score by classifying patients into 6 groups defined by the quintiles and the 90 percentile of the score in the combined cohort. The median survival months dramatically dropped from 192.6 in the 1st group (less than the 20th percentile) to 15.7 in the 6th group (above the 90th percentile). There appeared to exist a dose-response association: higher-percentile groups were associated with shorter survival and higher mortality risk (HR_{1 vs 6}=5.18, 95% CI: 3.73-8.45, \textit{P}=2.66 \times 10^{-6}; HR_{5 vs 6}=16.28, 95% CI: 10.57-25.07, \textit{P}=1.07 \times 10^{-16}; HR_{7 vs 6}=7.05, 95% CI: 4.66-10.69, \textit{P}=3.03 \times 10^{-3}; HR_{8 vs 6}=3.88, 95% CI: 2.51-6.00, \textit{P}=9.78 \times 10^{-4}; HR_{9 vs 6}=2.63, 95% CI: 1.69-4.10, \textit{P}=8.31 \times 10^{-3}; see Figure 2g-h).

**Predictive performance of APOLLO**

APOLLO predicted the 36- and 60-month survival rates quite accurately in the TCGA training set and CGGA1 internal testing set (AUC_{36-month}=0.933 and 0.888; AUC_{60-month}=0.854 and 0.851) (Figure 3a-b) and exhibited an excellent predictive ability in the CGGA2, Rembrandt, Weller and Gravendeel external testing sets (AUC_{36-month}=0.898, 0.893, 0.844 and 0.861, AUC_{60-month}=0.896, 0.817, 0.806 and 0.790) (Figure 3c-f). In the meta-analysis, APOLLO presented an excellent accuracy in both training sets (AUC_{36-month}=0.913, AUC_{60-month}=0.852) and testing sets (AUC_{36-month}=0.879, AUC_{60-month}=0.831), and combined data (AUC_{36-month}=0.901, AUC_{60-month}=0.843). The calibration curve suggested a good accordance (Supplementary Figure S3). APOLLO significantly outperformed a basic model with the four covariates aforementioned (Supplementary Figure S4), improving AUC by 5.4% (\textit{P}=2 \times 10^{-16}) and 5.8% (\textit{P}=2 \times 10^{-16}) for the 36- and 60-month survival prediction, respectively (Supplementary Figures S5-S6). Additionally, APOLLO presented an excellent C-index in the TCGA training cohort (0.874) and CGGA1 (0.804) internal testing cohort and four external testing cohorts: CGGA2 (0.870), Rembrandt (0.772), Weller (0.787) and Gravendeel (0.759); and a pooled C-index of 0.818 (95% CI: 0.800-0.835) (Figure 3i).

**Clinical net benefits with APOLLO**

With 36-month survival as the endpoint, DCA showed that APOLLO presented more clinical net benefits than several competing intervention strategies, namely, intervention for all, intervention for none, and intervention based on a basic model with only clinical and demographic indicators. Specifically, compared with the strategy of intervention for none and with a reasonable threshold probability (e.g., \textit{P}=0.4), APOLLO presented a higher net benefit (NB) than the basic model (NB_{\text{APOLLO}}=0.130 vs NB_{\text{Basic}}=0.111). In other words, APOLLO identified 13 true positive patients per 100 patients that ought to have intervention, whereas only 11.1 for the basic model (Figure 4a). On the other hand, compared to the strategy of intervention for all, APOLLO presented a higher net reduction (NR) than the basic model (NR_{\text{APOLLO}}=55.4% vs NR_{\text{Basic}}=52.5%). This means APOLLO can reduce the number of unnecessary clinical interventions by 55.4%, without missing interventions for any patients truly at high mortality risk; by comparison, only 52.5% for basic model (Figure 4b). As a sensitivity analysis and by varying the threshold probability from 0 to 0.5, the APOLLO decision curves were higher than those of the other strategies over a spectrum of threshold probability and APOLLO had the best average NB and NR in for 36- and 60-month survival (\text{NB}_{\text{36-month}}=0.166, \text{NR}_{\text{60-month}}=40.1\% and \text{NB}_{\text{60-month}}=0.258, \text{NR}_{\text{60-month}}=19.2\%), indicating its uniform utility and suitability for clinical implementation (Figure 4a-d). For individualized prognostic prediction and screening of high-risk patients, a nomogram of APOLLO is presented in Figure 4e.
Figure 2. Kaplan-Meier survival curves of LGG patients stratified by APOLLO score.

Survival differences between high- and low-risk patients in (A) TCGA, (B) CGGA1, (C) CGGA2, (D) Rembrandt, (E) Weller and (F) Gravendeel cohorts. Patients in all six cohorts were categorized into two groups based on the same cutoff point: the median of APOLLO score defined in TCGA training set. (G) Discriminative ability of the APOLLO score by illustrating the 36- and 60-month survival rate, median survival month for six groups, defined by quantiles at 20%, 40%, 60%, 80% and 90% of APOLLO score as the cutoffs. (H) The hazard ratios (HRs) and P values for patients at different levels of APOLLO score (level 1 as reference), which were derived from a Cox proportional hazards model.
Sensitivity analysis of APOLLO prediction

To assess the robustness of APOLLO, we performed a series of subgroup analyses with subgroups defined by age, gender, WHO grade, IDH mutation, 1p/19q status, MGMT promoter, radiotherapy and chemotherapy. In all the subpopulations examined, APOLLO presented good discriminative ability; the HRs that compare high- and low-risk groups within the subpopulations ranged from 3.33 (95% CI: 2.45-4.52, \( P=1.59 \times 10^{-14} \)) to 8.77 (95% CI: 5.65-13.61, \( P=4.54 \times 10^{-7} \)) (Supplementary Figure S7a). Moreover, APOLLO had reasonable AUCs in all of these subpopulations, ranging from 0.829 (95% CI: 0.784-0.873) to 0.907 (95% CI: 0.875-0.940) for 36-month survival and 0.757 (95% CI: 0.705-0.810) to 0.921 (95% CI: 0.881-0.961) for 60-month survival (Supplementary Figure S7b-c).

Figure 3. Time-dependent receiver operating characteristic curves of APOLLO for 36- and 60-month overall survival prediction.

The time-dependent ROC and AUC of APOLLO in (A) TCGA, (B) CGGA1, (C) CGGA2, (D) Rembrandt, (E) Weller and (F) Gravendeel cohorts, respectively. The pooled accuracy for (G) AUC\textsubscript{36-month}, (H) AUC\textsubscript{60-month} and (I) C-index of APOLLO across six independent cohorts.
Figure 4. Decision curve analysis and nomogram for clinical application of APOLLO.
In real-world applications, missingness may happen, in which case we recommend to use the mean imputation to fill the missing values of genes before applying APPOLO. Our simulations verified the feasibility of mean imputation (Supplementary Table S9).

Comparison of APPOLO with existing models by a systematic review
Among the 34 screened articles (Figure 5), the prognostic models have various types of predictors: 31 (57.4%) models were developed based on gene expressions, 8 (14.8%) on lncRNA and 6 (11.1%) on radiomic features (Supplementary Table S10). A total of 30 (55.6%) models were constructed by integrating multi-level biomarkers, and 19 (35.2%) studies considered molecular mutations. Except for 4 models that were only applicable to LGG subgroups (2 for IDH-wild type LGG, 1 for Grade II LGG and 1 for LGG with epilepsy), all of the models were suitable for all LGG patients. While differing in biomarker selection methods, 52 (96.3%) models were derived using Cox models. Of the 35 models using clinical variables, age was the most common predictor (n=34), followed by grade (n=27), IDH mutation (n=17), gender (n=8) and 1p/19q status (n=7) (Supplementary Table S10).

The prediction accuracy of these published LGG prognostic models was extracted from the original paper and was summarized in Table 1 and Supplementary Table S11. While 8 studies had sample size<1,000, the rest only has small to modest sample sizes, which may not guarantee the reliability of the prediction model. The 24 (44.4%) models without any self-reported external validation should be used with caution; though the other 30 (55.6%) models were externally validated, 7 of which were not completely externally validated, as they used the validation sets to screen the predictors. Further, only 4 models had multiple validations (Supplementary Table S11). In general, among 22 models that were validated by completely external testing sets, their prediction accuracy varies (C-index=0.753, Range: 0.620–0.830; AUC5-year=0.789, Range: 0.635–0.896 and AUC3-year=0.720, Range: 0.594–0.807) and was in general smaller than that of APPOLO derived from four external testing sets (C-index=0.780, Range: 0.759–0.807; AUC5-year=0.877, Range: 0.844–0.898 and AUC3-year=0.812, Range: 0.790–0.896).

Discussion
Wide variation exists in LGG survival, ranging from 1 to over 10 years,1–6 and patients at high risk of mortality may warrant close imaging monitoring and radical post-operative adjuvant therapy.7 Hence, there is an urgent need to develop accurate and robust prognostic prediction models for data-aided clinical decisions.8–10 Leveraging available public LGG transcriptome data from six independent cohorts, we adopted a 3-D analysis strategy to screen biomarkers and developed APPOLO. Derived from a large LGG cohort (TCGA) and validated in 5 trans-ethnicity cohorts with European and Asian populations, APPOLO exhibited an excellent prediction accuracy in the training and testing sets. Further, it offered good clinical net benefits for screening patients with high risk of mortality. Our systematic review also confirmed that APPOLO outperformed existing prediction models.

As the utility and transportability of prediction models can be affected by gaps between the training population and the target population that the model is applied to,16 we addressed this by proposing a 3-D analysis strategy, including Double types of effects, Double steps of screening and Double steps of modeling. The first one ensured the accuracy of the APPOLO by recognizing that G×G interactions which provided valuable insight into biological mechanisms of complex diseases.18,20 The latter two guaranteed the robustness of APPOLO. For example, our screening procedure identified biomarkers using a European population (TCGA) and validated those biomarkers using an Asian population (CGGA). This trans-ethnic validation revealed robustness of the transcriptional predictors. In the ensuing modeling procedures, APPOLO was trained using a TCGA cohort and was later applied to one internal and 4 external cohorts (CGGA1, CGGA2, Rembrandt, Weller and Gravendeel), and retained excellent prediction accuracy regardless of stratification by age, gender, WHO grade, IDH mutation, 1p/19q status, MGMT promoter methylation level, and history of radiotherapy or chemotherapy.

According to Global Burden of Disease, there are over 1.71 million brain & nervous system cancer patients worldwide,7 and 427.5 thousand (25%) are LGG. Assuming that LGG patients with probability of mortality ≥0.4 should be clinically intervened (Figure 5b and 5d), APPOLO yielded NR16-month=55.4% and NR60-month=52.4%, meaning that, compared to the most extreme strategy of offering interventions on every LGG patient, our model could help reduce 236.8 thousand (427.5 × 55.4%) and 138.5 thousand (427.5 × 32.4%) unnecessary interventions for short- and long-term survival outcome, respectively. In the future, APPOLO may, through customized biochips, offer maximized

The decision curve analysis for net benefit (NB) and net reduction (NR) of patients avoided unnecessary interventions at both 36-month (A–B) and 60-month (C–D) survival, respectively for APPOLO and the basic model composed of four common demographic and clinical predictors. (E) The nomogram for APPOLO. The value of each predictor can be converted into the corresponding points according to the axis in the top of nomogram. The sum of points for each predictor can correspond to the total points axis at the bottom of the nomogram and further used to estimate the patient’s 36- and 60-month survival rate.
benefits to patients and provide cost-effective precision medicine. As such, our manuscript may present a proof of concept.

We found that APOLLO outperformed 54 models we reviewed in prediction accuracy and robustness. Further, we briefly summarized the biological functions of these transcriptional biomarkers in APOLLO. For the genes with significant main effects, the genetic variants of CHIC2 are found in brain tumor tissues, and ITGA5 is a prognostic factor of gliomas; PLCG1 and IGF2BP2 are related to SUMOylation and m6A methylation, and involved in the immune responses, occurrence and development of gliomas; MSN is an active biomarker for glioma immune regulation and a drug target. For the pairs of genes with significant interactions, PRF1 is strongly associated with anti-CTLA-4 or anti-PD-L1 immunotherapy, and is related to immune cell activities and survival of gliomas. BCORL1 is a transcriptional corepressor that can fuse with ELF4, and repress the activation of PRF1. HMGA1 and TFG are regulated by NRF1, and can affect the prognosis of gliomas. FAS and SMAD4 are important members of the TNF-receptor superfamily and TGF-β signaling pathway, respectively, play a major role in tumor microenvironment and have antagonistic interactions. Though the biological function of the interaction between CTNND2 and GOLGA5 remains unclear, overexpressed CTNND2 is likely to increase tumor invasion of gliomas. PRF1, HMGA1, BCORL1, FAS, and MSN are the top 5 transcriptional biomarkers that are the most correlated with immune checkpoint genes. Specifically, PRF1 is viewed to be critically important for the immune cytolytic activity (CYT), reflecting the immune response of tumor cells, and is a well-established marker for cancer survival, including gliomas. HMGA1 contributes to the immuno-suppressive microenvironment in tumors and the silencing of HMGA1, and can boost checkpoint blockade immunotherapy. BCORL1 is involved in the immune response pathway, impacting the response to immunotherapy. FAS receptor signaling plays many important roles in the immune system, evidenced by that the tumoral FAS expression may predict the survival of CAR-T-treated patients. MSN, a known target for cancer immunotherapy, regulates the migration of effector T cells. Finally, genes included in APOLLO were transcriptional predictors with immune relevance, which can be immunotherapeutic targets.

Our study has several strengths. First, we performed, to our knowledge, the first systematic review of...
<table>
<thead>
<tr>
<th>No</th>
<th>PMID</th>
<th>Year</th>
<th>Method</th>
<th>Validation</th>
<th>Data source</th>
<th>Sample size</th>
<th>Performance in training set</th>
<th>Performance in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Training</td>
<td>Testing</td>
</tr>
<tr>
<td>1</td>
<td>33665000</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>505</td>
<td>143</td>
<td>1420</td>
</tr>
<tr>
<td>2</td>
<td>34123829</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>476</td>
<td>407</td>
<td>1114</td>
</tr>
<tr>
<td>3</td>
<td>33951297</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>506</td>
<td>592</td>
<td>1098</td>
</tr>
<tr>
<td>4</td>
<td>33400376</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>506</td>
<td>592</td>
<td>1098</td>
</tr>
<tr>
<td>5</td>
<td>33594759</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>495</td>
<td>590</td>
<td>1085</td>
</tr>
<tr>
<td>6</td>
<td>34395274</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>474</td>
<td>407</td>
<td>1049</td>
</tr>
<tr>
<td>7</td>
<td>33381460</td>
<td>2020</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>525</td>
<td>420</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>34015817</td>
<td>2021</td>
<td>NN</td>
<td>External</td>
<td>TCGA</td>
<td>493</td>
<td>408</td>
<td>901</td>
</tr>
<tr>
<td>9</td>
<td>33363544</td>
<td>2020</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>459</td>
<td>362</td>
<td>821</td>
</tr>
<tr>
<td>10</td>
<td>32519365</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>477</td>
<td>199</td>
<td>815</td>
</tr>
<tr>
<td>11</td>
<td>31824866</td>
<td>2019</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>511</td>
<td>172</td>
<td>683</td>
</tr>
<tr>
<td>12</td>
<td>31803233</td>
<td>2019</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>511</td>
<td>172</td>
<td>683</td>
</tr>
<tr>
<td>13</td>
<td>34408772</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>495</td>
<td>172</td>
<td>667</td>
</tr>
<tr>
<td>14</td>
<td>32793593</td>
<td>2020</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>476</td>
<td>170</td>
<td>646</td>
</tr>
<tr>
<td>15</td>
<td>31533943</td>
<td>2019</td>
<td>Cox</td>
<td>External</td>
<td>CGGA</td>
<td>172</td>
<td>451</td>
<td>623</td>
</tr>
<tr>
<td>16</td>
<td>31921517</td>
<td>2020</td>
<td>Cox</td>
<td>External</td>
<td>TCGA</td>
<td>456</td>
<td>159</td>
<td>615</td>
</tr>
<tr>
<td>17</td>
<td>24049111</td>
<td>2013</td>
<td>Cox</td>
<td>External</td>
<td>EORTC</td>
<td>338</td>
<td>235</td>
<td>573</td>
</tr>
</tbody>
</table>

Table 1 (Continued)
<table>
<thead>
<tr>
<th>No</th>
<th>PMID</th>
<th>Year</th>
<th>Method</th>
<th>ValidationType</th>
<th>Data source</th>
<th>Sample size</th>
<th>Performance in training set</th>
<th>Performance in testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Training</td>
<td>Testing</td>
<td>Training</td>
</tr>
<tr>
<td>18</td>
<td>29204839</td>
<td>2018</td>
<td>Cox</td>
<td>External</td>
<td>TCGA    CGGA</td>
<td>420</td>
<td>100</td>
<td>520</td>
</tr>
<tr>
<td>19</td>
<td>32162004</td>
<td>2020</td>
<td>RSF</td>
<td>External</td>
<td>Local   TCGA</td>
<td>205</td>
<td>91</td>
<td>296</td>
</tr>
<tr>
<td>20</td>
<td>30362964</td>
<td>2018</td>
<td>Cox</td>
<td>External</td>
<td>TCGA    CGGA</td>
<td>85</td>
<td>148</td>
<td>233</td>
</tr>
<tr>
<td>21</td>
<td>33409797</td>
<td>2021</td>
<td>Cox</td>
<td>External</td>
<td>Local   Local</td>
<td>149</td>
<td>66</td>
<td>215</td>
</tr>
<tr>
<td>22</td>
<td>32060714</td>
<td>2020</td>
<td>Cox</td>
<td>External</td>
<td>Local   TCGA</td>
<td>112</td>
<td>46</td>
<td>158</td>
</tr>
<tr>
<td>23</td>
<td>32740813</td>
<td>2020</td>
<td>Cox</td>
<td>External</td>
<td>Local   TCGA</td>
<td>117</td>
<td>33</td>
<td>150</td>
</tr>
<tr>
<td>24</td>
<td>32229719</td>
<td>2020</td>
<td>Cox</td>
<td>Internal</td>
<td>TCGA    CGGA</td>
<td>329</td>
<td>405</td>
<td>1216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GSE16011</td>
<td>140</td>
<td>118</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GSE61374</td>
<td>136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>31853837</td>
<td>2020</td>
<td>Cox</td>
<td>Internal</td>
<td>TCGA    CGGA</td>
<td>329</td>
<td>405</td>
<td>874</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GSEA1</td>
<td>140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>32351547</td>
<td>2020</td>
<td>Cox</td>
<td>Internal</td>
<td>TCGA    CGGA</td>
<td>304</td>
<td>353</td>
<td>785</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GSEA2</td>
<td>128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>32431729</td>
<td>2020</td>
<td>Cox</td>
<td>Internal</td>
<td>TCGA    CGGA</td>
<td>297</td>
<td>353</td>
<td>774</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GSEA2</td>
<td>124</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>33591634</td>
<td>2021</td>
<td>Cox</td>
<td>Internal</td>
<td>TCGA    CGGA</td>
<td>352</td>
<td>224</td>
<td>728</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GSEA2</td>
<td>152</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison of prediction accuracy between APOLLO and 28 models of LGG with self-reported external validation.

Abbreviations: NN: neural network; RSF: random survival forest; iAUC: integrated area under the time-dependent ROC curve; Internal: a model was cross validated by randomly splitting the original data. External: a model was externally validated by an independent external population. The performance for each model was extracted from the original paper.

\text{a} Datasets were used for biomarker screening, which were not completely external validation.
prognostic prediction models for LGG and confirmed the good performance of APOLLO. Second, this is perhaps the largest molecular prognostic prediction study for LGG, and APOLLO was strongly overall, as well as trans-ethnically validated by several large LGG cohorts. Our extensive subgroup analysis suggested the robustness and transportability of APOLLO to different populations. Third, we proposed an effective 3-D strategy for biomarker screening and model construction, by focusing on biomarkers with important main effects or G×G interactions. The strategy struck a reasonable balance among statistical properties (false positive control vs statistical power gaining), model interpretations (main effects vs G×G interactions), and computational complexity (fast variable screening vs consistent model selection). Finally, we provided a web-based tool to facilitate the application of APOLLO.

We also acknowledge some limitations. First, heterogeneity existed across these cohorts with various sequencing or microarray platforms. To address this, we harmonized the data by performing standard normal transformation, which work to some degree. Second, some well recognized prognostic factors (e.g., tumor size and extent of surgical resection) were missing in several cohorts. We envision that there is much room for improvement with more available and complete clinical factors. Third, applications of APOLLO to the other ethnicity populations should be cautious, as APOLLO was trained and validated among the Asian and European populations. Forth, the improvement of accuracy was not uniform in all external validation datasets, possibly due to the population heterogeneity or the limited sample size in a single dataset. Finally, more biological experiments are needed to confirm gene functions of these transcriptional predictors used in APOLLO.

To conclude, we presented an Accurate and independently validated Prediction moDel of Lower-grade gLio mas Overall survival (APOLLO), which was demonstrated, by a systematic review, with the best prediction accuracy and robustness, and was a cost-effective strategy for screening LGG patients at high risk of mortality. A free and user-friendly online tool was developed at http://bigdata.njmu.edu.cn/APOLLO.

Contributors
Study design: J.C., R.Z., Y.W., S.S. and F.C.; Data collection and quality control: J.C., J.F. and D.C.C.; Analyses and interpretation: J.C., J.F., Y.W., S.S., W.D., X.D., Y.Z., L.L., Y.L. and Z.L.; Online tool: J.C., C.Z.; Systematic review: J.C., J.F., X.Q. and L.L.; Manuscript draft: J.C., R.Z. and Y.L.; Manuscript revise: Y.L., Y.W., F.C., X.Q. and D.C.C.; All authors participated in TCGA, CGGA and GEO for providing the data. This study was supported by the National Key Research and Development Program of China (2016YFE0204900 to F.C.), National Science Foundation of Jiangsu Province (BK20191354 to R.Z.), National Natural Science Foundation of China (81973142 to Y.W. and 82103946 to S.S.), China Postdoctoral Science Foundation (2020M681671 to S.S.), the US National Institutes of Health (CA209414, CA249096, CA092824 and ES000002 to D.C.C., CA249096 and CA209414 to Y.L.), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). R.Z. was partially supported by the Qing Lan Project of the Higher Education Institutions of Jiangsu Province and the Outstanding Young Level Academic Leadership Training Program of Nanjing Medical University.

Supplementary materials
Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2022.104007.

References