Evaluating Dimensionality Reduction for Patient-Reported Outcome–Based Survival Modeling in Patients With Head and Neck Cancer

Eric Ababio Anyimadu, MS[1]; Yaohua Wang, PhD[1]; Amy C. Moreno, MS, MD[2]; Clifton David Fuller, MD, PhD[2]; Xinhua Zhang, PhD[3]; G. Elisabeta Marai, PhD[3]; and Guadalupe M. Canahuate, PhD[1]

DOI https://doi.org/10.1200/CCI-25-00069

ABSTRACT

PURPOSE This study aims to improve survival modeling in head and neck cancer (HNC) by integrating patient-reported outcomes (PROs) using dimensionality reduction techniques. PROs capture symptom severity across the treatment timeline and offer key insights for personalized care. However, their high dimensionality poses challenges such as overfitting and computational complexity. This work focuses on transforming and incorporating PRO data to enhance model performance in HNC.

MATERIALS AND METHODS We analyzed retrospective data of 923 patients with HNC treated at the University of Texas MD Anderson Cancer Center between 2010 and 2021. Baseline clinical data including demographic, treatment, and disease characteristics were used to build a reference survival model. PRO data, capturing symptom ratings, were integrated using dimensionality reduction techniques: principal component analysis (PCA), autoencoders (AEs), and patient clustering. These reduced representations, combined with clinical data, were input into Cox proportional hazards models to predict overall survival (OS) and progression-free survival (PFS). Model performance was assessed using the concordance index, time-dependent AUC, Brier score for calibration, and hazard ratios for predictor significance.

RESULTS Cox models incorporating PCA and AE outperformed the clinical-only reference model for both OS and PFS. The PCA-based model achieved the highest C-indices (0.74 for OS and 0.64 for PFS), followed by the AE model (0.73 and 0.63) and the clustering model (0.72 and 0.62). Time-dependent AUCs reinforced these results, with PCA showing the highest average AUC over 36 months. All models were well-calibrated, with low Brier scores. Key predictors included age, disease stage, and tumor subsite.

CONCLUSION Dimensionality reduction techniques improve survival prediction in patients with HNC by effectively incorporating PRO data, potentially providing greater insights into more personalized treatment strategies.

ACCOMPANYING CONTENT

Appendix

Data Sharing Statement

Accepted August 27, 2025
Published October 15, 2025

JCO Clin Cancer Inform
9:e2500069

INTRODUCTION

Head and neck cancers (HNCs) are malignant neoplasms arising in the nasal cavity, sinuses, lips, mouth, salivary glands, throat, or larynx. Despite advances in treatment, patients face high risk of recurrence or death.[1,2] According to the National Cancer Institute, an estimated 16,000 people in the United States are projected to die of HNC in 2024.[3]

Overall survival (OS) is the time from diagnosis or the start of treatment to death,[4] whereas progression-free survival (PFS) measures the time from treatment initiation to disease progression or death,[1] providing an earlier indicator of treatment efficacy. Both metrics inform treatment planning[5] and are central to oncology research.[6,7]

Identifying factors influencing OS and PFS is complex because of interactions among clinical characteristics, disease features, and treatment variability.[6] Traditional models rely on these variables,[5,8-10] but research shows that models integrating patient-reported outcomes (PROs) perform better. One study reported an increase in the concordance index (C-index) of OS models from 0.62 to 0.69 with PRO integration. PRO-based models also identified the American

CONTEXT

Key Objective

Can patient-reported outcomes (PROs) be effectively integrated into survival models for head and neck cancer (HNC) using dimensionality reduction techniques to improve predictive performance?

Knowledge Generated

This study shows that dimensionality reduction techniques such as principal component analysis and autoencoders effectively integrate PRO data into survival models, enhancing prediction accuracy in HNC. Models incorporating PROs outperformed those using clinical data alone.

Relevance *(F.P.-Y. Lin)*

PRO contain essential prognostic information, but their high-dimensional complexity has often limited their integration into clinical survival models. This study demonstrates that dimensionality reduction techniques can effectively harness this patient-centred data to create more accurate prognostic tools, enabling better decision-making and personalised treatment planning in the care of cancer patients.*

*Relevance section written by *JCO* CCI Deputy Editor Frank Po-Yen Lin, MBChB, PhD, FRACP, FAIDH.

Joint Committee on Cancer (AJCC) stage, age, and PROs as key OS predictors in patients with HNC.[8]

PROs are questionnaires that assess symptom severity before, during, and after treatment and offer valuable insights for personalized care and clinical decisions.[10] However, their high dimensionality, multicollinearity, and noise pose challenges for survival modeling, including overfitting and high computational complexity.[11,12] Robust transformation methods are needed to enhance their predictive utility. While some models use PRO-based clusters as survival predictors, this approach has limitations; symptom patterns evolve over time, and cluster assignments may vary across cohorts, reducing consistency.[8]

A promising approach for integrating PROs into survival models involves dimensionality reduction techniques such as principal component analysis (PCA) and autoencoders (AEs), which compress high-dimensional data into lower-dimensional forms, reducing noise while preserving key features.[13]

We evaluated survival models that incorporated low-dimensional PRO representations from PCA and AE alongside clinical variables. Additional models used PRO-based cluster labels as predictors, and a baseline model included only clinical variables. OS and PFS were used as outcomes.

Our results demonstrate that incorporating PROs through dimensionality reduction enhances predictive performance in survival modeling.

In summary, the contributions of this study are as follows:

- We propose methodology to integrate PRO into OS and PFS models using dimensionality reduction, specifically PCA and AE.
- We evaluate this method against the one that incorporates PROs through symptom severity clusters derived from patient stratification and another that relies solely on clinical variables without PROs.
- We show improved predictive performance and robustness in survival analysis compared with traditional approaches.

Related Work

Deep learning models, especially recurrent neural networks and their variants such as the long short-term memory (LSTM) and bidirectional LSTM, have been used to predict long-term symptom severity from previous PRO ratings.[8,14] Statistical methods, such as group-based trajectory modeling, also effectively capture heterogeneous symptom trajectories.[15] These approaches support the design of long-term personalized symptom management strategies in HNC.[8,14] However, the best approach to integrate PROs into clinical decision making remains largely unexplored.

Dimensionality reduction techniques, such as PCA and AE models, have been widely used in biomedical research to manage high dimensionality in data.[16]

PCA is a linear technique that reduces dimensionality by transforming data into orthogonal components while preserving variance.[13] It has been widely used in oncology, particularly for clinical and imaging data. In HNC, PCA has been applied to improve tumor detection in positron emission tomography scans of recurrent cases[17] and used to address multicollinearity in quality-of-life metrics to identify key survival-related components and hospitalization risks[18] and to analyze dose-volume histograms to reveal components linked to radiation-related complications.[19]

However, the linear nature of PCA may limit its ability to capture complex and nonlinear relationships that may exist in longitudinal PRO data.[20]

By contrast, AEs are neural networks that capture nonlinear patterns in high-dimensional data by learning compressed latent representations that retain essential information while reducing noise. AEs are increasingly used in oncology for disease classification, imaging, and survival prediction.[21] In HNC, AEs have been used to improve distant metastasis prediction by integrating multimodal radiomic features into unified representations[22] and enhanced organ-at-risk segmentation by generating latent features that boost accuracy, eliminate postprocessing, and support multiorgan segmentation.[23]

Despite their promise, PCA and AE have been underutilized for PRO data. We propose an approach that applies these techniques to compress longitudinal PROs into low-dimensional representations. These components are then integrated with clinical variables into OS and PFS prediction models.

MATERIALS AND METHODS

The following sections describe the methodology used in the study, with an overview provided in Figure 1.

Data

We analyzed data from 923 patients with HNC treated at the MD Anderson Cancer Center (2010-2021). Clinical variables were extracted from physician-completed records and included demographics (age, sex, smoking), disease or diagnostic characteristics (tumor subsite, T/N stage per AJCC 8th edition), and treatment indicators including induction chemotherapy (IC), concurrent chemotherapy (CC), and neck dissection surgery. Attributes with more than 99% homogeneity, for instance, radiation therapy receipt, M stage 0, and human papillomavirus–positive were excluded. Numerical variables such as age were min-max normalized to ensure uniform scaling.[24]

Categorical variables with <10% patient representation were merged for robustness. T stage was grouped into early (T0-T2) and late (T3, T4), and similarly N stage into N0, N1 and N2, N3. Tumor subsites other than base of tongue and tonsil were grouped into an "Other" category.

PROs were derived from patient self-reported symptom ratings which were collected using the MD Anderson Symptom Inventory (MDASI-HN questionnaire. Patients scored 28 symptoms from 0 ("not present") to 10 ("as bad as you can imagine"). Data were collected at five time points: before treatment (baseline), end of treatment, and post-treatment follow-ups at week 6, month 3, and month 12. MDASI-HN symptoms are grouped into general cancer, HNC-specific, and six interference items. However, this study focused on the 22 general and HNC-specific symptoms, excluding the interference symptoms because of their strong correlation with others, which could introduce redundancy and reduce model accuracy and interpretability.[8,18] Appendix Figure A1 presents the 22 symptoms analyzed in this study.

Finally, as with other longitudinal data sets based on patient responses, PROs often contain missing values because of incomplete responses, data capture issues, or dropout. Excluding patients with missing data can introduce bias as these individuals may differ systematically from the broader cohort, limiting generalizability.[25] To mitigate this, we included patients who rated all 22 symptoms at least once,

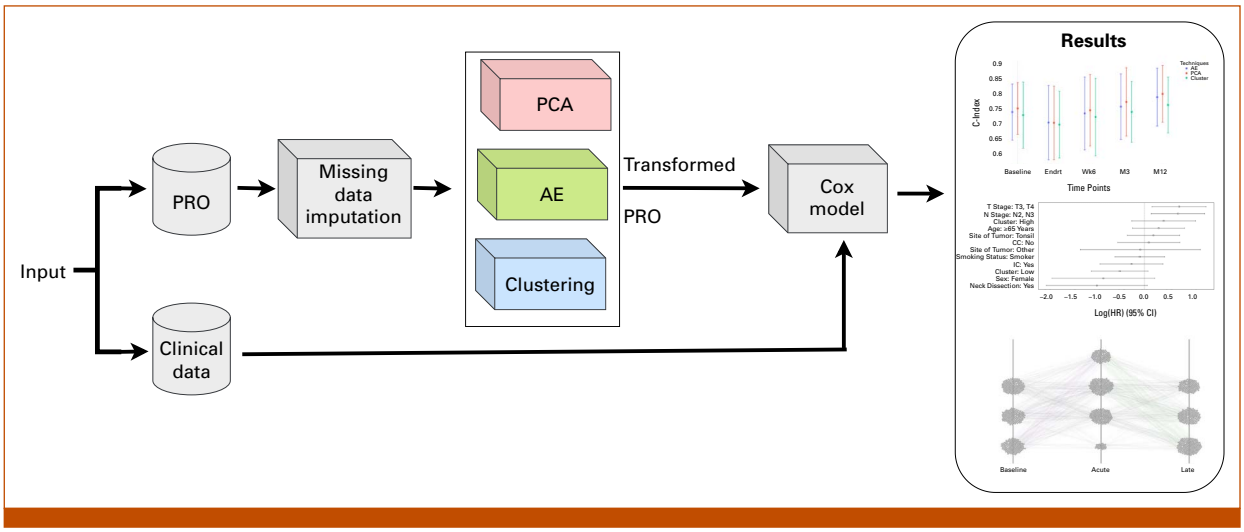

FIG 1. Overview of the study methodology, including data preprocessing, dimensionality reduction techniques (PCA and AE), clustering approach, Cox proportional hazards model (Cox model), and Results sections. AE, autoencoder; CC, concurrent chemotherapy; C-Index, concordance index; HR, hazard ratio; IC, induction chemotherapy; PCA, principal component analysis; PRO, patient-reported outcome.

even if they did not complete all time points, including those who died or dropped out before the end of follow-up.

To impute missing values, we used a symptom-based collaborative filtering method, which estimates missing PROs by leveraging intersymptom similarities via Pearson correlation.[26] This method has outperformed traditional approaches such as multiple imputation by chained equations, k-nearest neighbor, and linear interpolation in similar MDASI-HN data sets.[26,27]

This retrospective study is covered under the MD Anderson Institutional Review Board (IRB) protocol RCR-003-0800. In compliance with the Health Insurance Portability and Accountability Act, informed consent was waived and approved by the IRB as all analyses were performed over retrospective anonymized data.

Dimensionality Reduction

PCA and AE are used independently to reduce PRO data into principal and latent components, respectively, whereas clustering grouped patients into low and high PRO categories.

PCA

PCA used singular value decomposition to extract orthogonal components ranked by explained variance, enabling efficient dimensionality reduction by capturing the most informative features.[28]

AE

The AE architecture is an encoder-decoder design with a central bottleneck layer.[13] The encoder includes four dense layers: two with linear activations, one with rectified linear unit activation, and a final sigmoid layer. Input dimensions match the PRO data, and batch normalization between layers improves stability, reduces overfitting, and speeds up convergence.[29] The decoder mirrors the encoder's structure to effectively reconstruct the input from the latent space. The AE is trained to minimize mean squared error (MSE) between input and output using the Adam optimizer with a learning rate of 0.001 and a 20% internal validation to monitor performance. The complete AE architecture is shown in Appendix Figure A2.

Patient Clusters

The clustering of patients was performed based on their symptom ratings using hierarchical clustering, with Euclidean distance as the similarity measure and the Ward method as the linkage function, following the approach in a previous study.[8]

Survival Analysis

We used Cox proportional hazards models to evaluate the relationship between survival outcomes (OS/PFS) and predictor variables across several configurations: clinical-only, clinical plus PCA or AE components, and clinical plus symptom cluster labels.

Models were unregularized to allow clear interpretation of predictor effects. OS and PFS were modeled separately, with survival time in months. OS used a dead/alive event flag, whereas PFS included death or recurrence as events.

Evaluation

An 80/20 train-test split with five-fold cross-validation stratified by survival outcomes was used for robust evaluation. Missing values were imputed separately for train and test sets. To assess imputation accuracy, about 5% of observed symptom ratings were randomly held out and root mean squared error (RMSE) was computed between imputed and true values.[30]

To standardize PCA and AE transformations and reduce temporal variability, each symptom was treated as a single feature by aggregating its ratings across all time points and patients.

For both PCA and AE, models were fitted to the training set and applied to both training and testing sets for dimensionality reduction. A single principal component and one latent dimension were retained for PCA and AE for simplified interpretability, respectively. These representations were then used as inputs to the Cox model.

In the clustering approach, the training set was clustered first, and then the cluster labels were applied to classify patients in the testing set. This ensured consistent clustering and minimized instability from small test sets. Thus, patients in both sets were stratified into low and high symptom severity groups.

Dimensionality reduction can hinder interpretability by obscuring links between original features and reduced dimensions.[31] To clarify this, we derived stratifications from the transformed spaces, identifying two latent clusters and analyzing their symptom ratings. We also compared these with clusters from the original PRO data, providing insight into how the reduced dimensions align with original features.

The survival evaluation focused on data available at diagnosis, including clinical and baseline PRO. Models were assessed for discrimination and calibration. Discrimination was measured using the average C-index at 12 months post-treatment and the time-dependent AUC over 36 months (starting 12 months postdiagnosis, evaluated every 6 months). Calibration was assessed using the average Brier score at 12 months, measuring the difference between predicted and actual outcomes for event-free individuals. C-index, AUC, and Brier scores range from 0 to 1; higher C-index and AUC indicate better discrimination, whereas

lower Brier scores indicate better calibration.[32,33] Predictor significance in Cox models is reported as hazard ratios (HRs).

RESULTS

Data

Table 1 summarizes the clinical characteristics, survival outcomes, and baseline symptom clusters of the cohort of 923 patients. The median age was 60 years; 90.9% were male, and 43.55% were smokers. Most patients were diagnosed with tumor locations at the base of tongue or tonsils (90%), with 70% diagnosed at early T (T0-T3) and N (N0, N1) stages. Treatments received included CC (72.05%), IC (18.42%), and neck dissection (15.49%).

The median OS and PFS were both 35 months, with 8.56% deceased and 15.6% experiencing disease recurrence or death. Two baseline symptom burden clusters were identified: low (78.76%) and high (21.24%). Cluster distributions reflected the overall cohort. Significant associations were observed between clusters and T stage, OS, and PFS (P values < .001), with marginal associations for other variables (P value = .0069-.9730).

Considering the rate of missing values in the PRO data, the details on PRO completion by the patients are provided in Appendix Figure A3, which shows the number and proportion of patients with complete, partial, or missing symptom ratings at each time point. At baseline, 751 of 923 patients rated all 22 symptoms, 39 provided partial ratings, and 134 gave none. Week 6 had the greatest missingness, with only

TABLE 1. Patient Distribution of Clinical Data and Survival Outcomes for the Entire Cohort and Stratified Distribution by Baseline PRO-Based Cluster Labels

Attribute	Category	All Patients (N = 923, 100%), No. (%)	Low Symptom Cluster (n = 727, 78.76%), No. (%)	High Symptom Cluster (n = 196, 21.24%), No. (%)	*P*
Demographics					
Age, years	Median (25%-75%)	60 (54-67)	61 (54-67)	60 (55-66)	—
Sex	Male	839 (90.9)	671 (92.30)	168 (85.71)	.0069
	Female	84 (9.1)	56 (7.70)	28 (14.29)	
Smoking	Smoker	402 (43.55)	305 (41.95)	97 (49.49)	.0707
	Nonsmoker	521 (56.45)	422 (58.05)	99 (50.51)	
Disease specifics					
Site of tumor	BOT	427 (46.26)	335 (46.08)	92 (46.94)	.9730
	Tonsil	423 (45.83)	334 (45.94)	89 (45.41)	
	Other	73 (7.91)	58 (7.98)	15 (7.65)	
T stage	T0	56 (6.07)	50 (6.88)	6 (3.06)	<.001
	T1	300 (32.50)	252 (34.66)	48 (24.49)	
	T2	332 (35.97)	266 (36.59)	66 (33.67)	
	T3	140 (15.17)	101 (13.89)	39 (19.90)	
	T4	95 (10.29)	58 (7.98)	37 (18.88)	
N stage	N0	69 (7.48)	55 (7.57)	14 (7.14)	.0436
	N1	614 (66.52)	497 (68.37)	117 (59.69)	
	N2	217 (23.51)	156 (21.46)	61 (31.12)	
	N3	23 (2.49)	19 (2.6)	4 (2.04)	
Treatment specifics					
CC	Yes	665 (72.05)	520 (71.53)	145 (73.98)	.5556
	No	258 (27.95)	207 (28.47)	51 (26.02)	
IC	Yes	170 (18.42)	130 (17.88)	40 (20.41)	.4802
	No	753 (81.58)	597 (82.12)	156 (79.59)	
Neck dissection	Yes	143 (15.49)	122 (16.78)	21 (10.71)	.0486
	No	780 (84.51)	605 (83.22)	175 (89.29)	
Outcomes					
OS	Median (25%-75%)	35 (16-53)	35 (17-54)	32 (14-48)	—
	Alive	844 (91.44)	680 (93.35)	164 (83.67)	<.001
PFS	Median (25%-75%)	35 (16-52)	35 (17-54)	32 (14-48)	—
	Progression-free	779 (84.40)	630 (86.66)	149 (76.02)	<.001

Abbreviations: BOT, base of tongue; CC, concurrent chemotherapy; IC, induction chemotherapy; OS, overall survival; PFS, progression-free survival; PRO, patient-reported outcome.

479 patients providing complete responses and 403 providing none.

The pattern of missing data and the corresponding imputation performance across time points are summarized in Appendix Figure A4. The missing rate was about 15% at baseline, 35% at the end of treatment, and between 30% and 44% from week 6 to month 12. Imputation RMSE values were lowest at month 12 (1.563), month 3 (1.576), and baseline (1.589).

In terms of symptom severity, at baseline, pain, fatigue, sleep issues, and distress had the highest average severity ratings, whereas nausea, shortness of breath, vomiting, and skin symptoms were rated lowest. The details are provided in Appendix Figure A5.

Dimensionality Reduction

The distributions of PRO components derived from PCA and AE are examined to compare their patterns. Both sets of components were scaled to [0, 1] for easier comparison. The PCA components exhibit a broader spread, with most values clustered below 0.4, reflecting variance captured by the first principal component. By contrast, the AE components are more narrowly concentrated, with values tightly grouped below 0.1 or above 0.8. This supplementary comparison is illustrated using the swarm plot in Appendix Figure A6.

Figure 2 shows HNC-specific symptom severity grouped by clusters from PCA- and AE-transformed and PRO-based data. PCA and PRO clusters provided much clearer separation of low and high symptom burden levels than the AE clusters. Among the cohort, 159 were classified as high in the PCA clusters, 190 in the AE clusters, and 196 in the PRO clusters, showing similar overall compositions.

Consistent with overall severity patterns shown in Appendix Figure A5, symptoms such as dry mouth, mucus, swallowing, choking, voice, constipation, and mucositis remained relatively severe across both clusters, whereas skin and teeth symptoms were consistently mild.

Survival Analysis

Figures 3A and 3B shows average C-indices with 95% CIs for Cox models predicting OS and PFS. The reference model (clinical variables only) consistently had the lowest C-indices across training and testing. Adding PRO data improved discrimination. Among dimensionality reduction methods, PCA performed best (C-index: 0.74 OS, 0.64 PFS), followed by AE (0.73 OS, 0.63 PFS). Clustering by symptom severity also performed well (0.72 OS, 0.62 PFS).

Figure 4 shows time-dependent AUCs evaluated 12-36 months postdiagnosis at 6-month intervals for Cox models predicting OS. The PCA-based model had the highest average AUC (0.79), followed by AE and clustering. The reference had the lowest average AUC (0.72). Similar trends were observed for PFS.

Appendix Figures A7A and A7B shows average Brier scores with 95% CIs for OS and PFS, respectively. All models had low

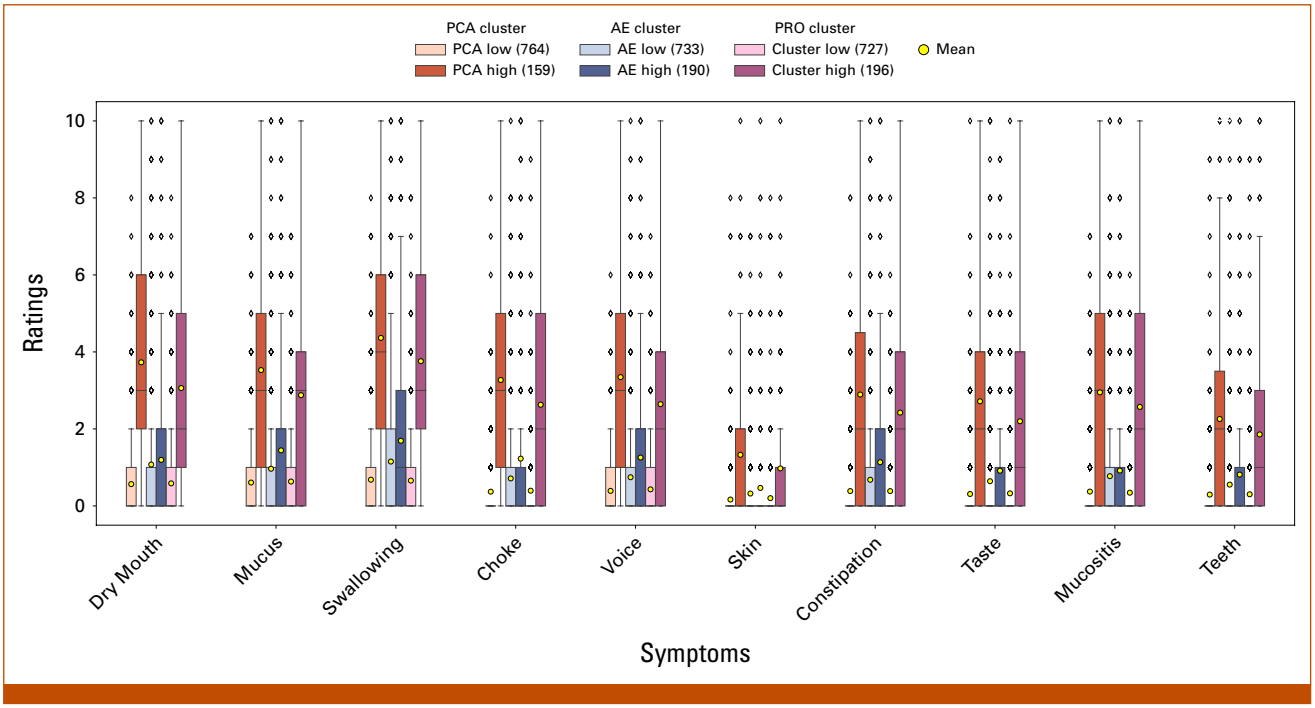

FIG 2. Box plot of stratified symptom severity: comparison across PCA, AE, and PRO-derived clusters for head and neck–specific symptoms. AE, autoencoder; PCA, principal component analysis; PRO, patient-reported outcome.

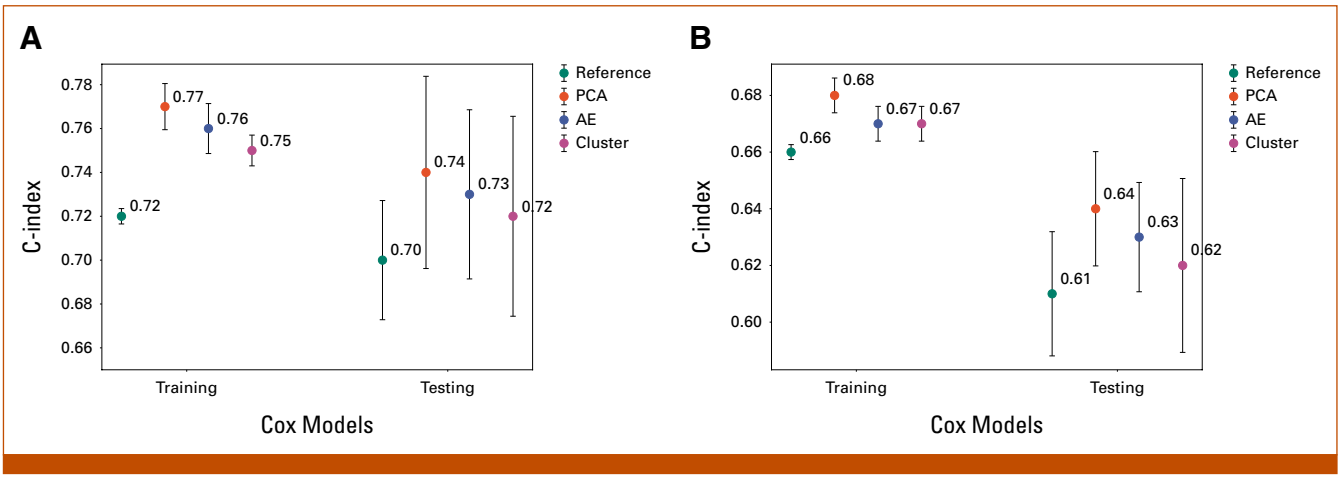

FIG 3. Average C-index with 95% CIs over five-fold cross-validated cox model evaluations for (A) OS and (B) PFS across training and testing sets. AE, autoencoder; C-index, concordance index; OS, overall survival; PCA, principal component analysis; PFS, progression-free survival.

scores, indicating good calibration. The clustering model slightly outperformed others for OS, whereas the reference model had a slightly better score for PFS.

Overall, the results underscore the value of incorporating PRO in survival models and show consistent training/testing trends, which suggest good generalization.

Figure 5A shows HRs for OS using the reference model, whereas Figure 5B shows hazard ratios for the model including the PRO PCA component. In the reference model, age, advanced T stage (T3, T4), and tonsil tumor site were associated with higher mortality, whereas female sex and neck dissection were associated with lower risk. With PCA included, the PRO PCA component was the strongest predictor of mortality. Similar trends are observed for PFS in Appendix Figures A8A and A8B.

DISCUSSION

This study demonstrates the value of integrating PRO into survival models for HNC. Using dimensionality reduction techniques such as PCA and AE and clustering, we combined PRO with clinical data for incorporation into survival models. PCA consistently outperformed others, achieving the highest C-indices for OS and PFS, suggesting that it captures key PRO patterns while reducing noise. AE showed competitive but slightly lower performance, indicating usefulness where

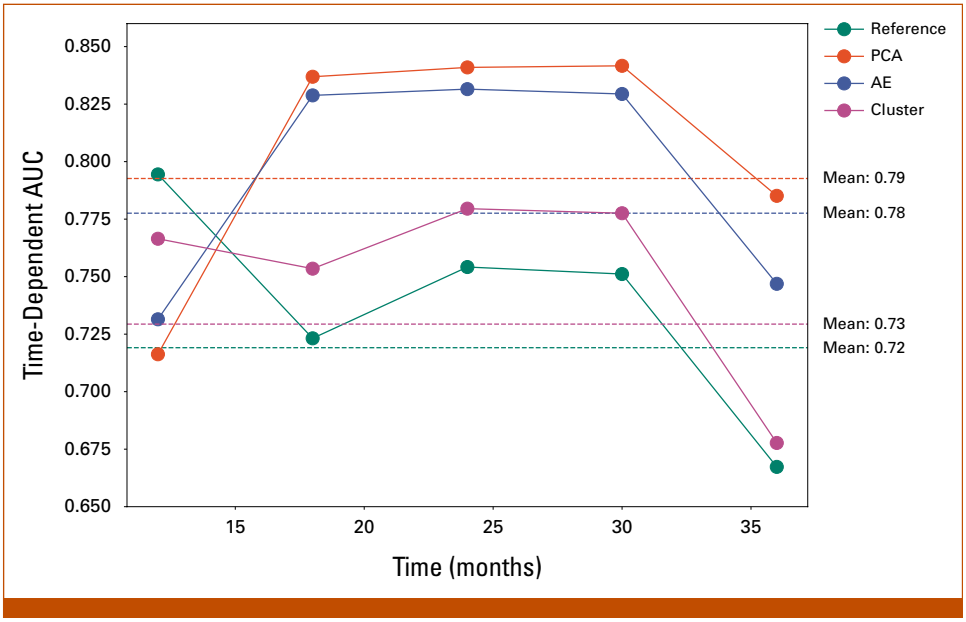

FIG 4. Time-dependent AUC evaluating Cox model performance over 36 months, assessed every 6 months starting from 12 months postdiagnosis for OS. AE, autoencoder; OS, overall survival; PCA, principal component analysis.

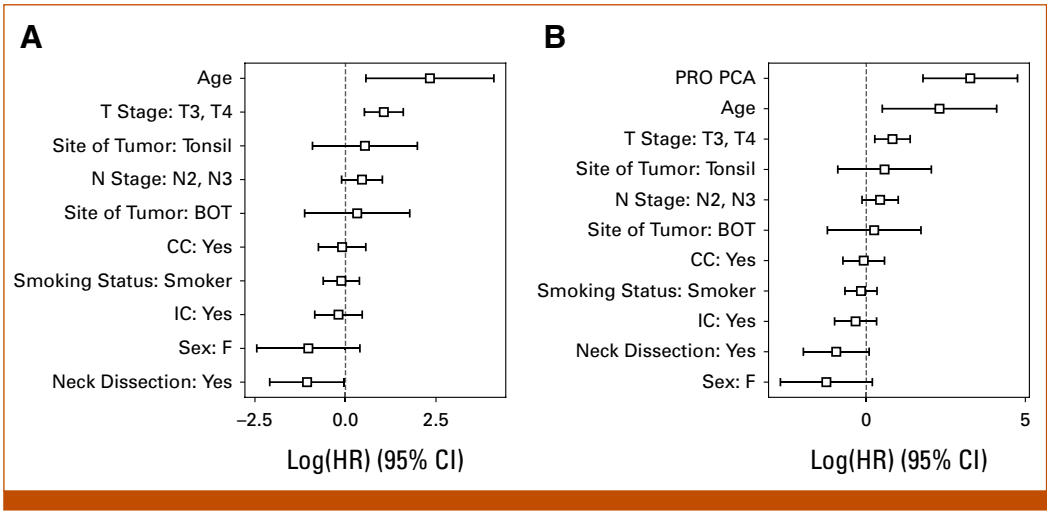

FIG 5. OS: comparison of HR for model predictors between (A) the reference model (clinical variables only) and (B) the model incorporating PRO data. BOT, base of tongue; CC, concurrent chemotherapy; HR, hazard ratio; IC, induction chemotherapy; OS, overall survival; PRO, patient-reported outcome.

nonlinear relationships matter. Clustering by symptom burden also provided reasonable predictive power, offering a practical alternative for clinical use.

In addition to the C-index, we assessed model discrimination over time using time-dependent AUC, measuring how well models distinguish patients with events from those who are event-free. The PCA-based model had the highest average AUC across follow-up, followed by AE and cluster models, confirming PCA's superior discrimination with PRO features.

Model calibration was assessed using the average Brier score at 12-36 months post-treatment. All models showed low scores, indicating good calibration. The clustering model performed best for OS, whereas the reference model had the lowest score for PFS. These results suggest that all models were well-calibrated despite differences in discrimination.

HR analysis showed PCA-transformed PRO components as key predictors of mortality and progression, outweighing clinical variables. This highlights the value of including PRO data in survival models. Remarkably, minimal components: one PCA component, one AE latent dimension, or two clusters, effectively captured PRO complexity and outperformed models without PRO data, demonstrating efficient dimensionality reduction without loss of predictive power.

A key challenge is the interpretability of transformed features, which can obscure relationships between original PRO variables and reduced components. To address this, we used cluster-based stratifications and analyzed symptom severity across clusters, revealing clearer alignment with original PRO data especially in PCA-derived clusters. This supports PCA's ability to capture meaningful PRO information. Although AEs can model more complex patterns, their training requires extensive hyperparameter tuning and architecture selection, complicating efficient model optimization.

The strong performance of PCA suggests that simple linear methods may be sufficient for modeling PRO data, aligning with previous findings on longitudinal PROs.[26]

The analysis of OS and PFS offered complementary perspectives on patient outcomes: OS captured overall mortality, whereas PFS served as an earlier indicator of disease progression and treatment efficacy. Clinical factors like age, tumor stage, and disease stage were consistently significant for both outcomes. Incorporating PRO data particularly via dimensionality reduction added nuance by highlighting symptom severity as a key survival predictor. These findings underscore the value of integrating PROs into clinical models to improve prognostication and support personalized treatment planning.

In survival risk modeling, dynamic prediction methods such as joint modeling and landmarking update time-to-event predictions using longitudinal data like PROs.[33,34] By contrast, our approach models survival using only PROs and clinical covariates available at diagnosis or before treatment. This enables early, proactive patient care with minimal data. While complementary to dynamic models, our framework highlights the utility of transformed baseline PROs within traditional survival models and provides a foundation for future integration of longitudinal PRO data.

In conclusion, this study highlights the value of integrating PROs into survival models, with PCA proving especially effective. Future work should further investigate the relationships between PROs, OS, and PFS to enhance and inform patient care.

AFFILIATIONS

[1]The University of Iowa, Iowa City, IA
[2]The University of Texas MD Anderson Cancer Center, Houston, TX
[3]The University of Illinois Chicago, Chicago, IL

CORRESPONDING AUTHOR

Guadalupe M. Canahuate, PhD; e-mail: guadalupecanahuate@uiowa.edu.

DISCLAIMER

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

SUPPORT

Supported in part by the National Institutes of Health (NIH) award (NCI-R01-CA258827; G.M.C.).

DATA SHARING STATEMENT

A data sharing statement provided by the authors is available with this article at DOI https://doi.org/10.1200/CCI-25-00069. The data supporting the findings of this study, including private patient data, are not publicly available due to restrictions, but may be obtained upon request from the corresponding author.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Financial support: Clifton David Fuller, G. Elisabeta Marai, Guadalupe M. Canahuate
Administrative support: Clifton David Fuller, G. Elisabeta Marai, Guadalupe M. Canahuate
Provision of study materials or patients: Amy C. Moreno, Clifton David Fuller
Collection and assembly of data: Eric Ababio Anyimadu, Yaohua Wang, Clifton David Fuller, Xinhua Zhang, G. Elisabeta Marai, Guadalupe M. Canahuate
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Clifton David Fuller
Employment: MD Anderson Cancer Center
Honoraria: Elekta, Varian Medical Systems/Siemens Healthineers
Research Funding: Elekta (Inst), RaySearch Laboratories (Inst), Oncospace (Inst)
Patents, Royalties, Other Intellectual Property: USPTO 11730561—Apparatus and methods for three-dimensional printed oral stents for head and neck radiotherapy [licensed to Kallisio Inc] (Inst)
Travel, Accommodations, Expenses: Elekta, Philips Healthcare, Siemens Medical Solutions USA, Inc
Other Relationship: NRG Oncology
Uncompensated Relationships: Philips/Elekta
Open Payments Link: https://openpaymentsdata.cms.gov/physician/444063

G. Elisabeta Marai
Patents, Royalties, Other Intellectual Property: Patent licensed to C-Motion Inc

No other potential conflicts of interest were reported.

REFERENCES

1. National Cancer Institute: Dictionary of cancer terms: Progression-free survival. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/progression-free-survival
2. Chow LQM: Head and neck cancer. N Engl J Med 382:60-72, 2020
3. American Association for Cancer Research (AACR): Head and Neck Cancers, 2024. https://www.aacr.org/patients-caregivers/cancer/head-and-neck-cancers/
4. National Cancer Institute: Dictionary of cancer terms: Overall survival. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/overall-survival
5. Mohamed AN, Wahid KA, Mohamed ASR, et al: Progression free survival prediction for head and neck cancer using deep learning based on clinical and PET/CT imaging data, in 3D Head and Neck Tumor Segmentation in PET/CT Challenge. Cham, Switzerland, Springer International, 2021, pp 287-299
6. Korn RL, Crowley JJ: Overview: Progression-free survival as an endpoint in clinical trials with solid tumors. Clin Cancer Res 19:2607-2612, 2013
7. Ganesan P, Sekaran S, Ramasamy P, et al: Systematic analysis of chemotherapy, immunotherapy, and combination therapy in Head and Neck Squamous Cell Carcinoma (HNSCC) clinical trials: Focusing on overall survival and progression-free survival outcomes. Oral Oncol Rep 12:100673, 2024
8. Anyimadu EA, Wang Y, Floricel C, et al: Pro-based stratification improves model prediction for toxicity and survival of head and neck cancer patients. IEEE J Biomed Health Inform 29:807-814, 2025
9. Canahuate G, Wentzel A, Mohamed ASR, et al: Spatially-aware clustering improves AJCC-8 risk stratification performance in oropharyngeal carcinomas. Oral Oncol 144:106460, 2023
10. Marai GE, Ma C, Burks AT, et al: Precision risk analysis of cancer therapy with interactive nomograms and survival plots. IEEE Trans Vis Comput Graph 25:1732-1745, 2019
11. Dormann CF, Elith J, Bacher S, et al: Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27-46, 2013
12. Witten DM, Tibshirani R: Survival analysis with high-dimensional covariates. Stat Methods Med Res 19:29-51, 2010
13. Wang Y, Yao H, Zhao S: Auto-encoder based dimensionality reduction. Neurocomputing 184:232-242, 2016
14. Wang Y, Van Dijk L, Mohamed ASR, et al: Improving prediction of late symptoms using LSTM and patient-reported outcomes for head and neck cancer patients, in 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), Houston, TX, IEEE, 2023, pp 292-300
15. Shi Q, Mendoza TR, Gunn GB, et al: Using group-based trajectory modeling to examine heterogeneity of symptom burden in patients with head and neck cancer undergoing aggressive non-surgical therapy. Qual Life Res 22:2331-2339, 2013
16. Franco EF, Rana P, Cruz A, et al: Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. Cancers 13:2013, 2021
17. Anzai Y, Minoshima S, Wolf GT, et al: Head and neck cancer: Detection of recurrence with three-dimensional principal components analysis at dynamic FDG PET. Radiology 212:285-290, 1999
18. Farrugia M, Yu H, Ma SJ, et al: A principal component of quality of life measures is associated with survival for head and neck cancer patients treated with radiation therapy. Cancers 13:1155, 2021

19. Dawson LA, Biersack M, Lockwood G, et al: Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation. Int J Radiat Oncol Biol Phys 62:829-837, 2005
20. Linting M, Meulman JJ, Groenen PJF, et al: Nonlinear principal components analysis: Introduction and application. Psychol Methods 12:336-358, 2007
21. Kabir MF, Chen T, Ludwig SA: A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthc Anal 3:100125, 2023
22. Zhou Z, Wang K, Folkert M, et al: Multifaceted radiomics for distant metastasis prediction in head & neck cancer. Phys Med Biol 65:155009, 2020
23. Tong N, Gou S, Yang S, et al: Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. Med Phys 45:4558-4567, 2018
24. Singh D, Singh B: Investigating the impact of data normalization on classification performance. Appl Soft Comput 97:105524, 2020
25. Weber GM, Adams WG, Bernstam EV, et al: Biases introduced by filtering electronic health records for patients with "complete data." J Am Med Inform Assoc 24:1134-1141, 2017
26. Anyimadu EA, Fuller CD, Zhang X, et al: Collaborative filtering for the imputation of patient reported outcomes, in Strauss C, Amagasa T, Manco G, et al (eds): Database and Expert Systems Applications. Lecture Notes in Computer Science, Volume 14910. Cham, Springer, 2024
27. Wang Y, Canahuate GM, Van Dijk LV, et al: Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes, in Proceedings of the 25th International Database Engineering & Applications Symposium (IDEAS '21). New York, NY, Association for Computing Machinery, 2021, pp 273-279
28. Salem N, Hussein S: Data dimensional reduction and principal components analysis. Procedia Comput Sci 163:292-299, 2019
29. Santurkar S, Tsipras D, Ilyas A, et al: How does batch normalization help optimization? Adv Neural Inf Process Syst 31, 2018
30. Hodson TO: Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geosci Model Dev 15:5481-5487, 2022
31. Chipman HA, Gu H: Interpretable dimension reduction. J Appl Stat 32:969-987, 2005
32. Steck H, Krishnapuram B, Dehing-Oberije C, et al: On ranking in survival analysis: Bounds on the concordance index. Adv Neural Inf Process Syst 20, 2007
33. Pickett KL, Suresh K, Campbell KR, et al: Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC Med Res Methodol 21:216-14, 2021
34. Putzel P, Smyth P, Yu J, et al: Dynamic survival analysis with individualized truncated parametric distributions, in Survival Prediction-Algorithms, Challenges and Applications. Proceedings of Machine Learning Research, 2021, pp 159-170

APPENDIX 1. SUPPLEMENTARY MATERIALS

Symptom Category

Appendix Figure A1 shows the MD Anderson Symptom Inventory general cancer- and head and neck cancer–specific symptoms considered in this study.

Autoencoder Architecture

The complete autoencoder (AE) architecture showing the linear, rectified linear unit and sigmoid layers with batch normalization layers and node dimensions is shown in Appendix Figure A2. The figure shows the optimal AE architecture chosen after grid search over the number and types of activation layers.

Patient-Reported Outcome Missing Rate and Imputation Performance

The details on patient-reported outcome (PRO) completion rates are provided in Appendix Figure A3, which presents the number and proportion of patients with complete, partial, or missing symptom ratings at each time point.

The pattern of missing data and the corresponding imputation performance across time points are shown in Appendix Figure A4.

Symptom Severity at Baseline

Appendix Figure A5 illustrates the distribution of symptom severity ratings at baseline.

Dimensionality Reduction

Appendix Figure A6 shows the distribution of principal component analysis (PCA)– and AE-transformed PRO data, both scaled to [0, 1] for comparability.

Brier Scores for Overall Survival and Progression-Free Survival

Appendix Figures A7A and A7B shows average Brier scores with 95% CIs for overall survival and progression-free survival (PFS), respectively.

PFS Model Hazard Ratios

Appendix Figure A8A shows hazard ratios (HRs) for PFS using the reference model, whereas Appendix Figure A8B shows HRs for the model including the PRO PCA component.

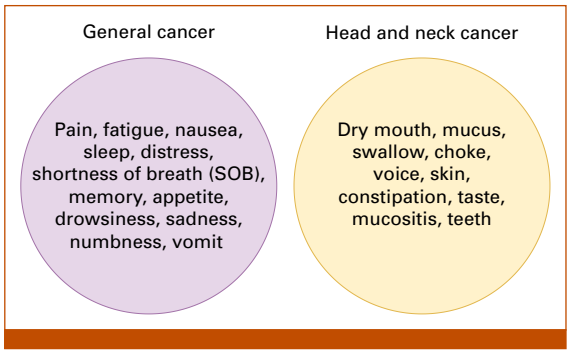

FIG A1. Symptoms from the MDASI-HN questionnaire categorized into general cancer symptoms and HNC-specific symptoms. HNC, head and neck cancer; MDASI-HN, MD Anderson Symptom Inventory; SOB, shortness of breath.

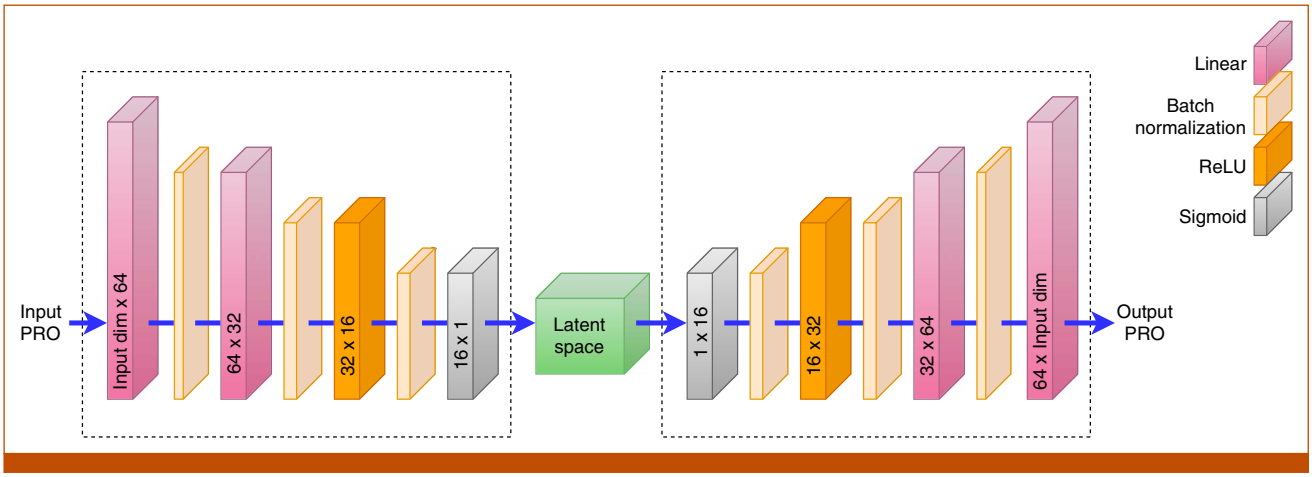

FIG A2. AE architecture for dimensionality reduction, featuring linear, ReLU, and sigmoid activation functions with batch normalization applied between activation layers. The node dimensions for both the encoder and the decoder are specified. AE, autoencoder; PRO, patient-reported outcome; ReLU, rectified linear unit.

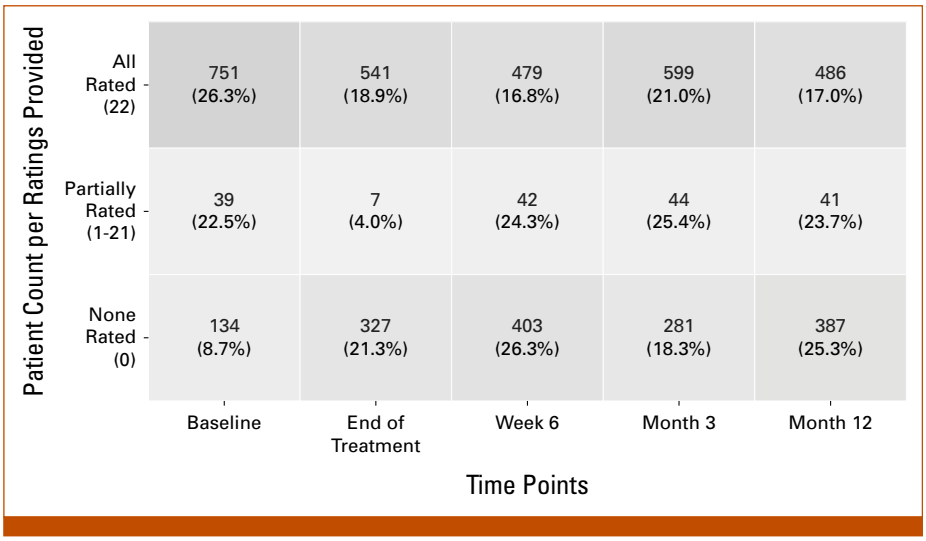

FIG A3. Counts and proportions of patients according to the number of ratings provided per time point.

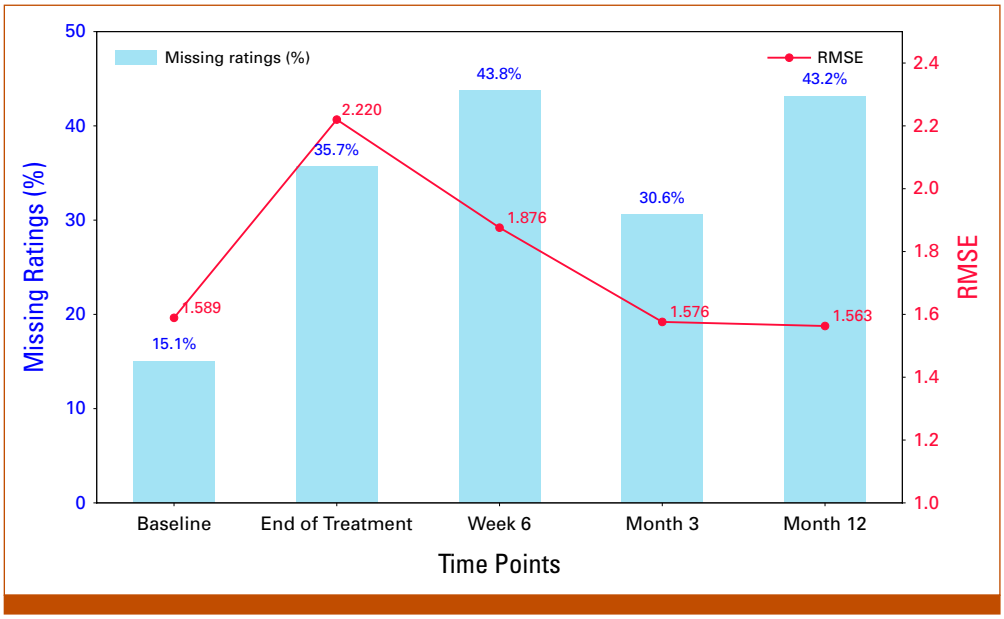

FIG A4. Missing ratings (%) and the RMSE imputation performance across the time points. RMSE, root mean squared error.

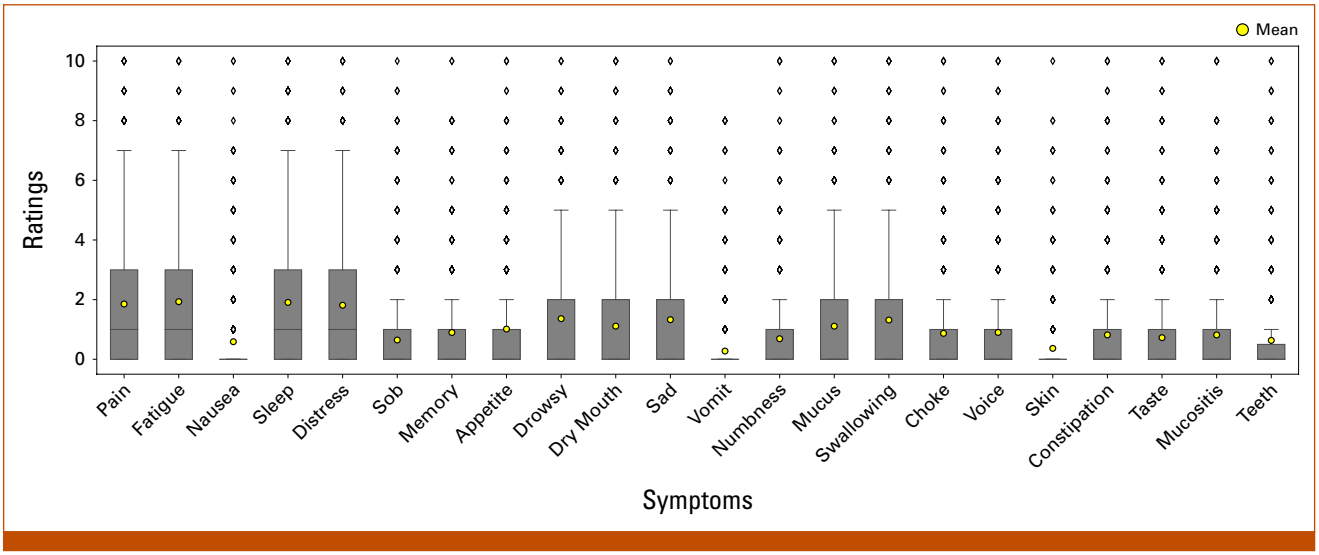

FIG A5. Box plot of symptom severity distribution across the data set.

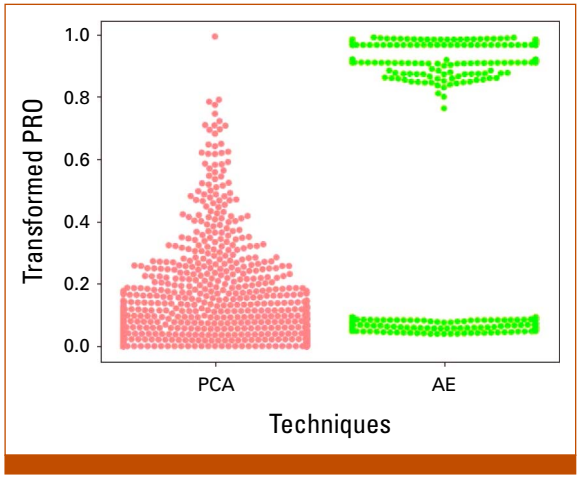

FIG A6. Swarm plot of PRO distribution in reduced space: PCA and AE techniques. AE, autoencoder; PCA, principal component analysis; PRO, patient-reported outcome.

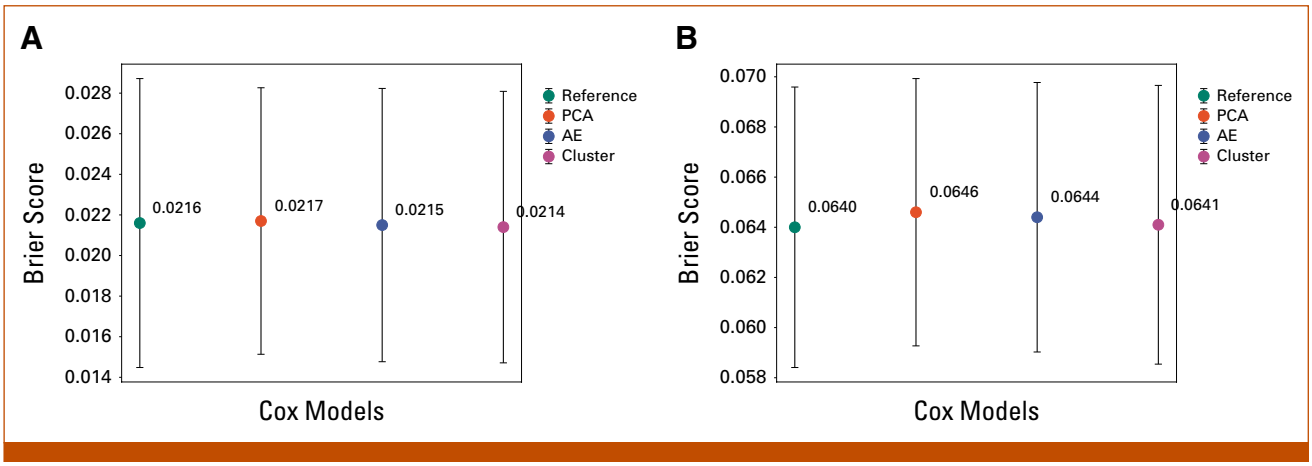

FIG A7. Average Brier score with 95% CIs over five-fold cross-validated Cox model evaluations for (A) OS and (B) PFS across the testing sets. AE, autoencoder; OS, overall survival; PCA, principal component analysis; PFS, progression-free survival.

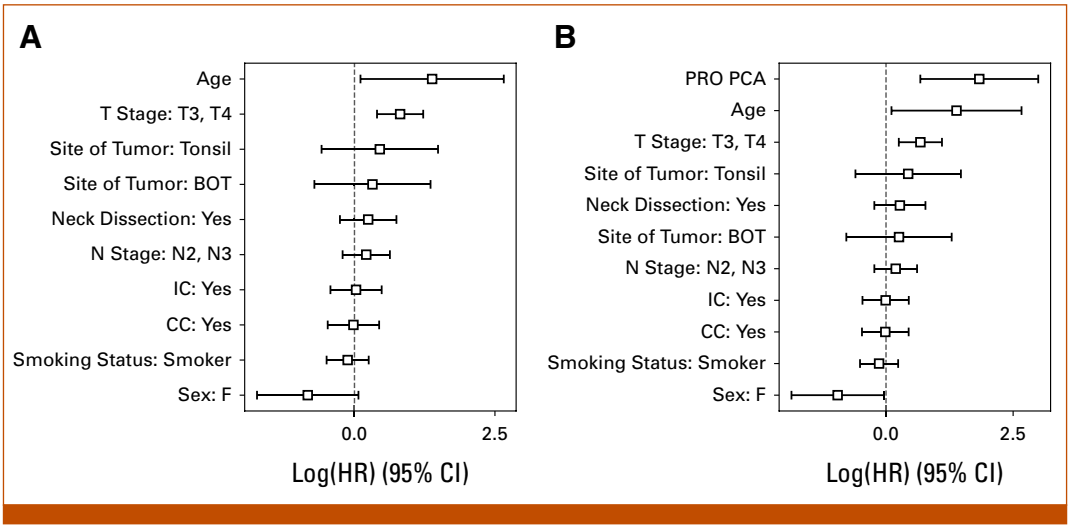

FIG A8. PFS: comparison of HR for model predictors between (A) the reference model (clinical variables only) and (B) the model incorporating PRO data. BOT, base of tongue; CC, concurrent chemotherapy; HR, hazard ratio; IC, induction chemotherapy; PFS, progression-free survival; PRO, patient-reported outcome.