Joint probability framework for the development and validation of a prognostic model for the conditional outcome of quality of life: a retrospective study in historical European cohorts of survivors of head and neck cancer

Mauricio Moreira-Soares, Erlend I F Fossen, Aritz Bilbao-Jayo, Aitor Almeida, Laura Lopez-Perez, Itziar Alonso, Maria Fernanda Cabrera-Umpierrez, Giuseppe Fico, Susanne Singer, Katherine J Taylor, Steve Thomas, Miranda Pring, Lisa Licitra, Stefano Cavalieri, Arnoldo Frigessi, Marissa LeBlanc

Summary

Background Conditional outcomes are outcomes defined only under specific circumstances. For example, future quality of life (QoL) can only be ascertained when patients are alive. In prognostic models involving conditional outcomes, a choice must be made on the precise target of prediction: one could target future QoL, given that the individual is still alive (conditional) or target future QoL jointly with the event of being alive (unconditional). We aimed to (1) introduce a probabilistic framework for prognostic models for conditional outcomes, and (2) apply this framework to develop a prognostic model for QoL 3 years after diagnosis in patients with head and neck cancer.

Methods A joint probability framework was proposed for prognostic model development for a conditional outcome dependent on a post-baseline variable. The framework involved two submodels: one for QoL and one for survival. Joint probability was estimated with conformal estimators using MAPIE with least absolute shrinkage and selection operator (LASSO) regression and XGBoost models. We included patients with head and neck cancer who were alive with no evidence of disease 12 months after diagnosis from the UK-based Head and Neck 5000 cohort (N=3572) and made QoL predictions 3 years after diagnosis. Predictors included clinical and demographic characteristics and longitudinal measurements of QoL. External validation was performed in longitudinal studies led by the University of Mainz (Mainz, Germany) and the Istituto Nazionale dei Tumori (Milan, Italy). A total of 497 patients were used in external validation of the QoL submodel and 281 were used in external validation of the survival submodel. Model performance was evaluated with C-statistics for discrimination, calibration plots, and R^2 or mean absolute error (MAE), or both, for overall performance as appropriate.

Findings Of 3572 patients, 400 (11·2%) were dead by the time of prediction (3 years after diagnosis), whereas 73 (26·0%) of 281 patients were dead in the validation set for the survival submodel. Model performance was assessed for prediction of QoL, both conditionally and jointly with survival. C-statistics ranged from 0·66 to 0·74 in internal validation and 0·60 to 0·80 in external validation. In internal validation, R^2 and MAEs ranged from 0·37 to 0·5 and 12·2 to 13·5, respectively, whereas R^2 and MAEs ranged from 0.35 to 0.41 and 13.0 to 12.3, respectively, in external validation. The calibration plots showed reasonable calibration in external validation. External performance was weaker for the survival submodel than for the QoL submodel. An application programming interface and dashboard were developed.

Interpretation Our probabilistic framework for conditional outcomes provides both joint and conditional predictions and thus the flexibility needed to answer different clinical questions. Our model had reasonable performance in external validation and has potential as a tool in long-term follow-up of QoL in patients with head and neck cancer.

Funding EU's Horizon 2020 research and innovation programme.

Lancet Digit Health 2026; 8: 100976

Published Online April 15, 2026 https://doi.org/10.1016/j.landig.2025.100976

Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway (M Moreira-Soares PhD, E I F Fossen PhD, Prof A Frigessi MSc, M LeBlanc PhD); **CICERO Center for International Climate Research, Oslo, Norway** (E I F Fossen); **DeustoTech, Faculty of Engineering, University of Deusto, Bilbao, Bizkaia, Spain** (A Bilbao-Jayo PhD, A Almeida PhD); **Universidad Politécnica de Madrid-Life Supporting Technologies Research Group, ETSIT, Madrid, Spain** (L Lopez-Perez PhD, I Alonso MSc, M Fernanda Cabrera-Umpierrez PhD, G Fico PhD); **Institute of Medical Biostatistics, Epidemiology, and Informatics, University Medical Centre of Johannes Gutenberg University Mainz, Mainz, Germany** (S Singer PhD, K J Taylor PhD); **University Cancer Center at University Medical Centre of Johannes Gutenberg University Mainz, Mainz, Germany** (S Singer); **Comprehensive Cancer Center Mecklenburg-Vorpommern (CCC-MV), Department of Quality of Life in Oncology, University Medical Centre Rostock, Rostock, Germany** (S Singer); **Bristol Dental School, Bristol, United Kingdom** (S Thomas PhD, M Pring PhD); **Head and Neck Medical Oncology Department, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy** (L Licitra MD, S Cavalieri MD); **Department of Oncology and Hemato-oncology, University of Milan,**

Introduction

Conditional outcomes can only be measured under specific conditions. The focus is on outcomes that happen in the future, together with a conditioning event that also happens in the future. For example, future health-related quality of life (HRQoL) is only measurable when a person is still alive. In these situations, two variables are considered, both happening in the future. In prognostic models involving conditional outcomes, a choice must be made on the precise aim of the prediction: either one is interested in the conditional outcome or in the unconditional (joint) one. For the HRQoL example, the former targets quality of life, given that the individual is alive (conditional); the latter targets quality of life jointly with the event of still being alive (unconditional). The predictive probability of the first case assumes that the person is alive; in the second case the life status of the individual is also evaluated as a random event that might occur or not. Emphasis is frequently placed on

Milan, Italy (L Licitra, S Cavalieri); **Department of Method Development and Analytics, Norwegian Institute of Public Health, Oslo, Norway** (M LeBlanc)

Correspondence to: Dr Marissa LeBlanc, Oslo Centre for Biostatistics and Epidemiology, University of Oslo, 0317 Oslo, Norway **marissaerin.leblanc@fhi.no**

Research in context

Evidence before this study

We searched PubMed for studies published from database inception up to Sept 19, 2024, using the search terms "head and neck" AND "quality of life" AND ("prognostic prediction" OR "machine learning" OR "prediction model") and found 45 results. The prognostic models developed in the identified publications either excluded patients who died during follow-up or imputed quality of life with zero for patients who died during follow-up. None of these publications explicitly addressed the implications of conditioning on survival, which introduces a substantial risk of bias and could lead to invalid interpretations. These issues are well known in biostatistics and epidemiology but are often overlooked among machine learning practitioners and data scientists working with health data. Furthermore, methodological studies have been raising awareness about the importance of predicting probabilities that are well calibrated and suitable for answering the predictive questions of interest.

A systematic review published in 2019 showed that health-related quality of life in survivors of head and neck cancer can be severely impaired even 10 years after treatment. A 2021 scoping review highlighted the need for prediction models that support quality of life in survivors of cancer. Of the 67 studies included in the review, 33 (49%) conducted parametric tests, 32 (48%) used regression models to identify prognostic factors, and only two (3%) applied survival analysis and a non-linear method.

Added value of this study

This study makes an important methodological contribution that can be broadly applied to prognostic modelling in patient populations with non-negligible mortality, for whom survival is not the main target of prediction. To our knowledge, this is the first study to address this problem in the context of clinical prognostic models using a statistically sound approach. In addition, our proposed solution is model-agnostic and suitable for modern machine learning applications.

The study also makes an important clinical contribution for long-term follow-up (3 years after diagnosis) of patients with head and neck cancer by developing a joint prognostic model for quality of life and survival. To our knowledge, our model is the first joint model of long-term quality of life and survival in this patient population, with internal and external validation in European longitudinal studies of patients with head and neck cancer.

Implications of all the available evidence

The probabilistic framework proposed can affect the future development of clinical prediction models by raising awareness and proposing a solution for a ubiquitous problem in the field. The joint model can be tailored to address different clinical needs, for example to identify patients who are both likely to survive and have low quality of life in the future, or to predict an individual patient's future quality of life, either conditional or unconditional on survival. The model should be validated further in patient data from different countries.

the conditional outcome,[1] leading to predictions that condition on the occurrence of the post-baseline event. We argue that relying solely on conditional probability often does not yield sufficiently informative predictions for decision makers, who could be interested in the unconditional event. Because both occur in the future, it can be relevant to assess the probability that the outcome occurs, while incorporating the probability that the conditioning event also happens, instead of assuming this to be true. A common choice in prognostic prediction is to develop the model only in individuals for whom the conditional outcome is observed, often without acknowledging the conditionality.[2,3] Some studies attempt to include all patients by imputing the conditional outcome for the censored cases,[4] which introduces a substantial risk of bias and could lead to invalid interpretations.

In clinical prognostic models, the aim is to answer relevant clinical questions, ideally in a probabilistic manner. Previous prognostic models for HRQoL in patients with cancer have used survival to the time of prediction as an inclusion criterion,[5] and therefore estimate the conditional probability. This approach restricts the clinical question to "what is the probability of having a certain quality of life in the future, assuming the patient survives?" The alternative question is "what is the probability of both surviving and having a certain quality of life in the future?" This alternative question incorporates survival as a random event and considers that some patients are unlikely to survive in the first place. Estimating the joint probability of surviving and having an HRQoL level within a specified range therefore appears relevant.

Although our study has a generic value, we concretely focus on patients with head and neck cancer, who often endure a substantial symptom burden because of the disease or the treatment, or both, including swallowing, speaking, and breathing difficulties,[6] eating problems,[7] pain,[8] difficulties with social contacts,[9] and loss of income.[10] Typically, patients with head and neck cancer have a decline in HRQoL following the start of treatment compared with at diagnosis, yet they tend to have improved HRQoL after treatment conclusion, usually within 1 year,[11–13] although impairments can remain a problem for up to 10 years or longer.[14] It is therefore of clinical interest to develop a prognostic model for HRQoL over the long term (ie, 2 years following the end of treatment, equivalent to 3 years after diagnosis in our study). Such predictions could help clinicians to identify patients at high risk for HRQoL decline and implement early interventions to mitigate this risk.

This study aims to (1) introduce a probabilistic framework for prognostic models for conditional outcomes that

depend on post-baseline variables, with flexibility to address different clinical questions, and (2) apply this framework to develop and validate a joint prognostic model for HRQoL and survival over the long term in patients with head and neck cancer.

Methods

Probabilistic framework

For each patient $i \in \{1,\ldots,N\}$, let $\boldsymbol{x_i} = (\boldsymbol{x_{i1}},\ldots,\boldsymbol{x_{ip}})$ be a random vector of p predictors measured at a specific baseline timepoint t_0, and $y_i \in \mathbb{R}$ be a continuous outcome measured at a follow-up timepoint t_f, where $t_0 < t_f$. We defined the post-baseline event with the survival function given in equation 1:

$$S(t \mid \boldsymbol{X}) = P(T > t \mid \boldsymbol{X}) \text{ for all } t \geq 0,$$

that is, the probability that the time to event (T) is greater than some specific time t. The measurement of y at time $T=t$ is conditionally dependent on survival to time t. We then defined the joint probability of the continuous outcome being between the lower and upper thresholds (h_L, h_U) and survival to time t in equation 2:

$$P(h_L \leq y \leq h_U, T > t \mid \boldsymbol{X}) = P(h_L \leq y \leq h_U \mid \boldsymbol{X}, T > t) \\ P(T > t \mid \boldsymbol{X}),$$

in which the first term in the right-hand side of the equation is the conditional probability of measuring the continuous outcome between the thresholds given the survival time $T > t$, and the second term is the survival probability (this can be defined more generally by considering that the outcome y can belong to any given set G $[y \in G]$ but for simplicity we defined $G = h_L \leq y \leq h_U$ in equation 2, which is particular to the case example we explored in this paper.) To estimate the joint probability, we developed two submodels for (1) the conditional outcome in equation 3:

$$P(h_L \leq y \leq h_U \mid \boldsymbol{X}, T > t) = P(y \leq h_U \mid \boldsymbol{X}, T > t) - P(y \leq h_L \mid \boldsymbol{X}, T > t)$$

in which we estimate the probability $P(y \leq h \mid \boldsymbol{X}, T > t)$ for a generic threshold h so at inference time we can retrieve the probability in the desired interval by providing the specific lower and upper boundaries, and (2) the model for estimating the probability for time to event: $P(T > t \mid \boldsymbol{X})$.

Probability estimation

To obtain the probabilities in equation 3, we used the model agnostic prediction interval estimator[15] with cross-validation plus, which is a flexible estimator for predictive distributions of continuous outcomes. For submodel (2), any time-to-event method can be used, but the data in our case example present no censoring so we used least absolute shrinkage and selection operator (LASSO) logistic regression to estimate the survival probability to time t=2 years after the end of treatment.

Application of the framework to BD4QoL

The big data models and intelligent tools for QoL monitoring and participatory empowerment of survivors of head and neck cancer (BD4QoL) project (grant number 875192) was funded by the European Commission with partners from seven European countries. In this project, several datasets of patients with head and neck cancer were available and are described throughout.

Data sources

The Head and Neck 5000 study (HN5000) is a UK-based study of individuals with head and neck cancer and was used as the development set. It is sponsored by the University Hospitals Bristol and Weston National Health Service (NHS) Foundation Trust and conducted in collaboration with the University of Bristol.[16,17] From April 7, 2011, to Dec 31, 2014, 5511 participants were recruited at diagnosis from 76 head and neck centres in the UK. In the BD4QoL project, we had access to 143 quality-controlled variables collected from diagnosis up to 3 years of follow-up, encompassing clinical, demographic, HRQoL, physical, and mental health data. Survival data were obtained from participating sites and from NHS England. The HN5000 study was approved by the National Research Ethics Committee (South West Frenchay Ethics Committee, reference number 10/H0107/57, on Nov 5, 2010) and the Research and Development departments of participating NHS Trusts. Written informed consent was obtained from all included patients.

External validation was performed in longitudinal studies including patients with head and neck cancer at diagnosis, led by the University of Mainz and the Istituto Nazionale dei Tumori, with recruitment spanning from 2001 to 2017 (appendix pp 4–5 for study details). The Italian study was approved by the institutional research ethics board at the Istituto Nazionale dei Tumori with identifier number INT 65/16. The German studies received ethical approval from either the Leipzig University Ethics Committee or by Landesärztekammer Rheinland-Pfalz ethics committee under approval number 367 837.281.14 (9520). Written informed consent was obtained from all patients recruited to all studies.

See Online for appendix

The BD4QoL study protocol received ethical approval by the Norwegian Regional Committee for Medical and Health Research Ethics South-East D under application number 154191.

All studies collected HRQoL data with the European Organisation for Research and Treatment of Cancer Core Questionnaire (EORTC QLQ-C30) and its head and neck cancer module (EORTC QLQ-H&N35).[18,19]

We screened for eligibility criteria (appendix p 1) to identify patients who were alive and had no evidence of disease at 12 months after diagnosis (the first visit after end

of treatment), resulting in 3572 participants who were eligible for model development.

Outcome

The outcome is defined based on the Global Health Status/Quality of Life (GHS/QoL) measured with the respective scale in EORTC QLQ-C30[18] at 3 years after diagnosis. We will refer to GHS/QoL as simply QoL. The QoL is a pseudocontinuous scale that can take on eight possible values between 0 and 100, derived from two questions.[20]

Using the proposed probabilistic framework we can make prognostic predictions for QoL. Clinically relevant questions could include "what is the probability that, 2 years from now, this patient is alive and with lower QoL than today?" or "if this patient survives the next 2 years, what is the probability that their QoL is above a certain value?"

We show how to use the framework to model one particular target (ie, the probability of being alive with a future QoL less than or equal to a given threshold). Here the time of prediction t_f is at 2 years after the end of treatment (ie, 3 years after diagnosis in the HN5000 dataset). In the HN5000 dataset, we had complete information for survival (ie, no censoring because all patients were observed until 3 years after diagnosis or until death). Therefore, instead of modelling survival as a time to event, we modelled it as binary variable $S(t{=}t_f) := \{0, 1\}$ defined at the time of prediction t_f that indicates if the patient was alive at prediction time or not. This made the presentation easier, but more general situations are straightforward.

Applying our probabilistic framework, we defined the outcome measurement as the QoL score $y : QoL$. Thus, equation 3 was developed into equation 4:

$$P\left(QoL_L \leq QoL \leq QoL_U, S \,\middle|\, \boldsymbol{X}\right) = P\left(QoL_L \leq QoL \leq QoL_U \,\middle|\, \boldsymbol{X}, S = 1\right) P(S = 1|\boldsymbol{X}).$$

For the **configuration files** see https://github.com/ocbe-uio/BD4QoL

In this application, clinically significant deterioration in QoL was of interest, defined as a decrease of 10 points or more between t_f and t_0, and referred to as QoL decline throughout. Therefore, we will obtain the QoL decline probability after the general model estimation for the probability in equation 4 by defining $\text{QoL}_L := 0$ and $\text{QoL}_U := \text{QoL}_{t_0} - 10$. Equation 4 was thus developed as equation 5:

For the **dashboard** see https://ocbe-uio.github.io/bd4predict/

$$P\left(QoL \leq QoL_{t_0}{-}10, S \,\middle|\, \boldsymbol{X}\right) = P\left(QoL \leq QoL_{t_0}{-}10 \,\middle|\, \boldsymbol{X}, S = 1\right) P(S = 1|\boldsymbol{X}),$$

or, in terms of the complement, the probability of having a QoL decline or dying, $D{=}S^C$, was developed into equation 6:

$$P\left(QoL \leq QoL_{t_0}{-}10 \cup D \,\middle|\, X\right) = P\left(QoL \leq QoL_{t_0}{-}10 \,\middle|\, X, D = 0\right)(1 - P(D|X)) + P(D|X).$$

See the appendix (pp 1–2) for complete derivation.

Candidate predictors

We developed models with two different sets of predictors: (1) an extensive set of 143 variables (171 parameters) to identify prognostic factors in an agnostic manner and (2) a restricted subset of 43 variables (59 parameters).[21]

The extensive set included all available predictors in the data measured at diagnosis, at 4 months after diagnosis, and at the end of treatment at 12 months (ie, measured at or before time 0 in our model). The restricted set was chosen due to the variables' known association with QoL[22] and their routine availability in the clinical setting. These variables included demographic information, clinical data, QoL questionnaires (eg, EORTC QLQ-C30 and its module EORTC QLQ-H&N35[18]), and other psychometric information. Descriptions of the two lists of predictors can be found in the appendix (p 8).

Sample size

The estimated sample sizes[23] required for developing the prognostic models are 1430 and 4249 participants for 59 and 171 predictor parameters, respectively. The number of available participants was 1631, which is adequate for the restricted set but could increase the risk of overfitting for the extended set. To address this issue, we used an agnostic penalisation method. The details of this calculation as well as the sample size calculation for external validation can be found in the appendix (p 2).

Statistical analysis

We used Apptainer with Python version 3.10 for model development and Conda environment for exploratory analysis. Descriptive statistics were performed in R version 4.2.3. Median and IQR are shown for continuous characteristics. Counts and percentages are shown for the categorical characteristics. The configuration files containing the full list of packages with versions and the code are available online. We used Shapley Additive Explanations (SHAP) to estimate feature importance and the top 1% of mean absolute SHAP interaction values. We deployed models with an application programming interface (API) developed with FastAPI and a dashboard in HTML and JavaScript. This study adhered to the TRIPOD+AI reporting guideline.[24]

Missing data

Missing values in continuous and ordinal predictors were imputed with the k-nearest neighbour (KNN) method (k=5 and uniform weights). For categorical variables, the missing indicator method was used. We did not impute missing outcomes but a propensity scores model was developed for evaluating model performance (appendix pp 2–3). We only imputed variables that presented less than 50% missingness; the remaining variables were excluded from the analysis. To address potential biases in the model performance metric due to missing outcomes, we developed a propensity score model for

missingness. As an extra analysis in the development set, we stratified the patients by high and low propensity, to see if the C-statistic differed.

Missing variables in the validation set were handled by the model pipeline; for categorical variables the prediction model included a missing indicator and for continuous variables the KNN method was applied.

Model development

We modelled the QoL score distribution with multivariable conformal models,[25,26] to obtain the probability of QoL decline. The base estimators were LASSO linear regression and boosted trees with XGBoost. The predicted risk of death was modelled with LASSO logistic regression. Further details can be found in the appendix (p 3).

Internal and external validation

In the development set, we performed bootstrap internal validation with 200 resamples, applying the .632+ estimator to obtain the C-statistics, R^2, and the mean absolute error (MAE).[27] We used the .632 estimator[28] to obtain the observed-to-expected ratio, calibration-in-the-large, and calibration slope. Calibration plots were estimated with locally estimated scatterplot smoothing. See the appendix (p 3) for a complete description. External validation for the conformal model was performed with independent data in the validation sets.

Patient and public involvement

Although patients were not directly involved in the development of the prognostic model, the broader BD4QoL project has incorporated patient and public involvement through shared activities such as patient association workshops and online seminars to ensure that patient perspectives informed the overall study design and dissemination strategy.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

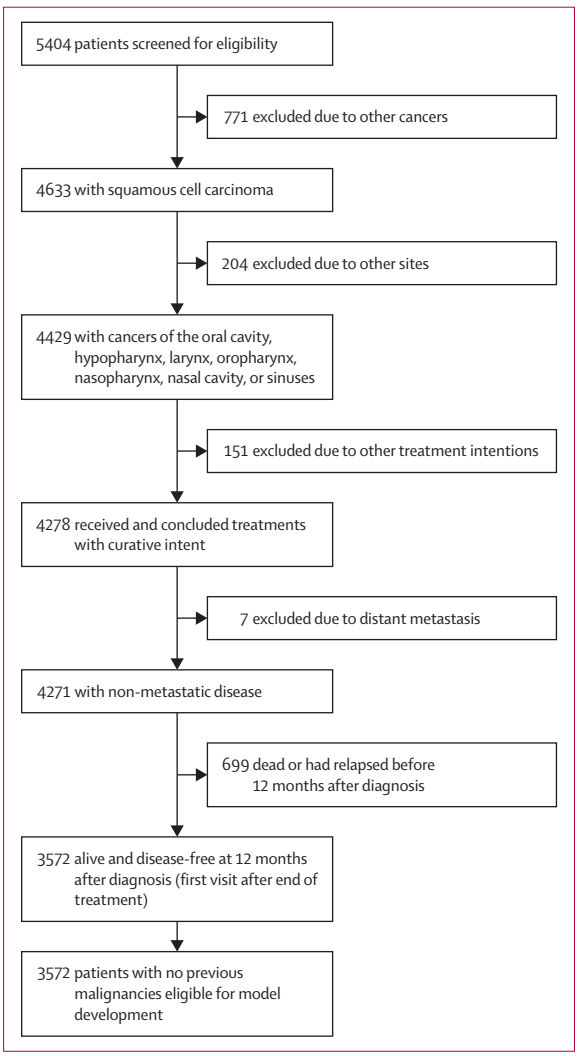

Figure 1: **Flow of participants in the study according to inclusion and exclusion criteria (appendix p 1)**

The Head and Neck 5000 cohort was screened to identify patients alive and disease-free 12 months after diagnosis. The percentages refer to the percentage of total screened.

Results

After screening the development data for the eligibility criteria, 3572 participants from HN5000 were included in the analyses (figure 1). Survival status was available for all participants, and 400 (11·2%) were dead at the time of prediction (3 years after diagnosis). The QoL outcome was measured in 1662 (52·4%) of 3172 remaining alive participants. After screening validation data with the eligibility criteria, a total of 497 patients were available with measured QoL and 281 patients were available for validation of the survival submodel, and 73 patients (26·0%) were dead at the time of prediction.

Baseline characteristics for the development set are presented in table 1. Participants who were alive at the end of treatment but dead 3 years after diagnosis were more often male, older, had higher cancer stages, more comorbidities, a higher prevalence of human papillomavirus (HPV)-negative tumours, lower income, lower education level, and higher levels of smoking and alcohol consumption than patients who were alive 3 years after diagnosis. Of the patients who were alive 3 years after diagnosis, those missing QoL were more often male, had more comorbidities, lower household income, were not married, had a lower education level, lived in more deprived areas, and smoked more compared with those with measured QoL (appendix pp 9–11).

Baseline characteristics for the external validation sets are presented in the appendix (pp 15–18). The major difference between the validation and development cohorts was in the distribution of tumour location, with 844 (24%) of

	Total (N=3572)	Alive at follow-up (N=3172)	Dead at follow-up (N=400)
Sex			
Male	2694 (75·4%)	2361 (74·4%)	333 (83·2%)
Female	878 (24·6%)	811 (25·6%)	67 (16·8%)
Age	61 (54–68)	61 (54–67)	65 (59–72)
TNM stage, UICC TNM Atlas, 7th edition			
I	851 (23·8%)	793 (25·0%)	58 (14·5%)
II	661 (18·5%)	579 (18·3%)	82 (20·5%)
III	497 (13·9%)	437 (13·8%)	60 (15·0%)
IV	1554 (43·5%)	1354 (42·7%)	200 (50·0%)
Missing	9 (0·3%)	9 (0·3%)	0
Tumour location			
Oral cavity	971 (27·2%)	852 (26·9%)	119 (29·8%)
Oropharynx	1509 (42·2%)	1376 (43·4%)	133 (33·2%)
Nasopharynx	69 (1·9%)	62 (2·0%)	7 (1·8%)
Hypopharynx	135 (3·8%)	104 (3·3%)	31 (7·8%)
Larynx	844 (23·6%)	741 (23·4%)	103 (25·8%)
Nasal cavity or sinuses, or both	44 (1·2%)	37 (1·2%)	7 (1·8%)
Treatment received			
Surgery	1700 (47·6%)	1516 (47·8%)	184 (46·0%)
Chemotherapy	1510 (42·3%)	1343 (42·3%)	167 (41·8%)
Radiotherapy	2748 (76·9%)	2408 (75·9%)	340 (85·0%)
Comorbidities, ACE-27			
No comorbidity	1550 (43·4%)	1439 (45·4%)	111 (27·8%)
Mild	1204 (33·7%)	1084 (34·2%)	120 (30·0%)
Moderate	615 (17·2%)	496 (15·6%)	119 (29·8%)
Severe	134 (3·8%)	93 (2·9%)	41 (10·2%)
Missing	69 (1·9%)	60 (1·9%)	9 (2·2%)
HPV status			
Not obtained	2415 (67·6%)	2121 (66·9%)	294 (73·5%)
Negative	284 (8·0%)	237 (7·5%)	47 (11·8%)
Positive	873 (24·4%)	814 (25·7%)	59 (14·8%)
BMI	26 (23–29)	26 (23–30)	25 (22–28)
Missing	1064 (29·8%)	915 (28·9%)	149 (37·3%)
QoL at baseline	75 (58–83)	75 (58–83)	66 (45–83)
Missing	1444 (40·4%)	1235 (38·9%)	209 (52·3%)
Household income			
Less than £3999	105 (2·9%)	87 (2·7%)	18 (4·5%)
£4000–7999	262 (7·3%)	230 (7·3%)	32 (8·0%)
£8000–11 999	316 (8·8%)	272 (8·6%)	44 (11·0%)
£12 000–17 999	380 (10·6%)	329 (10·4%)	51 (12·8%)
£18 000–22 999	223 (6·2%)	207 (6·5%)	16 (4·0%)
£23 000–28 999	257 (7·2%)	237 (7·5%)	20 (5·0%)
£29 000–34 999	188 (5·3%)	168 (5·3%)	20 (5·0%)
£35 000 or more	581 (16·3%)	549 (17·3%)	32 (8·0%)
Missing	1260 (35·3%)	1093 (34·5%)	167 (41·8%)
Marital status			
Single	331 (9·3%)	286 (9·0%)	45 (11·2%)
Widowed	177 (5·0%)	148 (4·7%)	29 (7·2%)
Separated	62 (1·7%)	55 (1·7%)	7 (1·8%)
Divorced	265 (7·4%)	229 (7·2%)	36 (9·0%)
Married	1533 (42·9%)	1397 (44·0%)	136 (34·0%)
Living with a partner	297 (8·3%)	277 (8·7%)	20 (5·0%)
Missing	907 (25·4%)	780 (24·6%)	127 (31·8%)

(Table 1 continues on next page)

	Total (N=3572)	Alive at follow-up (N=3172)	Dead at follow-up (N=400)
(Continued from previous page)			
Education level			
Primary education	33 (0·9%)	27 (0·9%)	6 (1·5%)
Secondary education	1164 (32·6%)	1033 (32·6%)	131 (32·8%)
Post-secondary education	886 (24·8%)	803 (25·3%)	83 (20·8%)
Higher education	450 (12·6%)	415 (13·1%)	35 (8·8%)
Missing	1039 (29·1%)	894 (28·2%)	145 (36·2%)
Deprivation, IMD10			
Least deprived	568 (15·9%)	516 (16·3%)	52 (13·0%)
Less	592 (16·6%)	538 (17·0%)	54 (13·5%)
Middle	698 (19·5%)	613 (19·3%)	85 (21·2%)
More	597 (16·7%)	532 (16·8%)	65 (16·2%)
Most deprived	652 (18·3%)	551 (17·4%)	101 (25·2%)
Missing	465 (13·0%)	422 (13·3%)	43 (10·8%)
Smoker status			
Never	577 (16·2%)	546 (17·2%)	31 (7·8%)
Former	1487 (41·6%)	1350 (42·6%)	137 (34·2%)
Current smoker	500 (14·0%)	404 (12·7%)	96 (24·0%)
Missing	1008 (28·2%)	872 (27·5%)	136 (34·0%)
Alcohol consumption, units per week	22 (9–43)	21 (9–36)	34 (14–50)
Missing	1558 (32·4%)	1362 (42·9%)	196 (49·0%)

Data are n (%) or median (IQR). Education levels are grouped as: primary education (ages 5–11 years), secondary education (ages 11–16 years), post-secondary education (ages 16 years and older; includes sixth form, further education, and vocational or technical courses), and higher education. Comorbidity was measured with the ACE-27. The IMD10 is a UK-specific deprivation index. ACE-27=Adult Comorbidity Index. HPV=human papillomavirus. IMD10=Index of Multiple Deprivation 2010. QoL=quality of life. UICC=Union for International Cancer Control.

Table 1: **Descriptive statistics for the included cohort, stratified by vital status at prediction time (3 years after diagnosis)**

3572 patients with tumours in the larynx in the development set against 334 (67%) of 497 patients in the validation set. In addition, the validation sets did not collect BMI, regional deprivation index, comorbidity index, and HPV status. Overall, there were no considerable differences for other clinical and demographic characteristics (table 1 *vs* appendix pp 15–18).

We developed conformal LASSO and XGBoost models for the continuous QoL outcome (n=1662), using either the restricted or extended set of predictors, and estimated the predictive distribution. For predicting mortality at 3 years after diagnosis (no censoring), we used logistic regression with L1-penalty, with 3572 patients and 400 events. The conformal LASSO and XGBoost models were used to predict the probability of QoL decline (equation 5), with 238 QoL declines in 1351 patients. The combined models for predicting QoL decline or death were then applied to 1751 observations and 638 events (appendix p 11).

The full list of estimated parameters is given in the appendix (pp 11–15). To use our prediction model, one can either directly interface with our API or use our dashboard, the BD4QoLPredict tool. Users input demographic, clinical, and QoL data, and the model predicts the joint probability of survival and QoL decline, estimated QoL score with prediction intervals, and the predictive conformal distribution, imputing missing values as needed. Imputed features are displayed in the dashboard for further analysis (appendix p 21). Complete API documentation and dashboard guidelines are provided in the respective repositories.

Table 2 presents the performance metrics for the joint models and submodels developed with LASSO and XGBoost. Overall, models developed with the restricted set of predictors presented similar discrimination (C-statistic 0·73 *vs* 0·72 for XGBoost and 0·66 *vs* 0·68 for LASSO) and slightly better calibration for the LASSO survival model with an observed-to-expected ratio of 1·13 versus 1·54, calibration-in-the-large of –0·15 versus –0·34, and calibration slope of 0·90 versus 0·81, and for the joint model with an observed-to-expected ratio of 1·19 versus 1·26, calibration-in-the-large of 0·11 versus 0·05, and calibration slope of 0·84 versus 0·78, compared with the extensive counterpart. All models had a moderate degree of miscalibration in different directions. Considering the usability and interpretability of simpler models and the close performance and better calibration, LASSO with the restricted set of predictors is the recommended model to be used in practice. We evaluated this model in two groups of patients that presented high and low propensity scores for missing outcome and found similar performance in patients resembling patients with missing outcomes versus patients not resembling patients with missing outcomes (appendix p 15).

Figure 2 presents the calibration plots and histograms for (A) continuous QoL conditional on survival, (B) the

For the **API** see https://bd4predict.azurewebsites.net

For the **BDQoLPredict tool** see https://ocbe-uio.github.io/bd4predict

		Conformal models for QoL					
		QoL decline $P(Decline \mid X, \text{Death} = 0)$			Continuous QoL		
Method	Predictors	C-statistic (95% CI)	Observed-to-expected ratio (95% CI)	Calibration-in-the-large (95% CI)	Calibration slope (95% CI)	R^2 (95% CI)	MAE (95% CI)
XGBoost	Restricted	0·69 (0·62 to 0·75)	1·49 (0·25 to 11·3)	0·25 (0·04 to 0·44)	2·53 (2·44 to 2·62)	0·44 (0·37 to 0·51)	12·91 (12·01 to 13·95)
XGBoost	Extensive	0·72 (0·66 to 0·79)	1·48 (0·27 to 11·3)	–0·26 (–0·45 to –0·06)	1·26 (1·18 to 1·35)	0·50 (0·45 to 0·55)	12·21 (11·46 to 13·02)
LASSO	Restricted	0·66 (0·61 to 0·69)	0·83 (0·28 to 5·14)	–0·72 (–0·98 to –0·41)	0·69 (0·51 to 0·90)	0·37 (0·32 to 0·41)	13·50 (13·05 to 14·02)
LASSO	Extensive	0·68 (0·64 to 0·72)	0·88 (0·28 to 6·05)	–0·74 (–0·99 to –0·44)	0·69 (0·53 to 0·86)	0·41 (0·36 to 0·46)	13·08 (12·53 to 13·73)

		Survival model: $P(Death \mid X)$			
		C-statistic (95% CI)	Observed-to-expected ratio (95% CI)	CIL (95% CI)	Calibration slope (95% CI)
Logistic regression	Restricted	0·74 (0·71 to 0·76)	1·13 (0·39 to 8·91)	–0·15 (–0·44 to 0·16)	0·90 (0·78 to 1·06)
Logistic regression	Extensive	0·72 (0·69 to 0·75)	1·54 (0·37 to 11·7)	–0·34 (–0·61 to –0·01)	0·81 (0·69 to 0·98)

		Joint model: $P(Decline \cup Death \mid X)$			
		C-statistic (95% CI)	Observed-to-expected ratio (95% CI)	CIL (95% CI)	Calibration slope (95% CI)
XGBoost	Restricted	0·73 (0·69 to 0·76)	1·50 (0·39 to 6·90)	0·33 (0·20 to 0·51)	1·13 (1·04 to 1·22)
XGBoost	Extensive	0·72 (0·68 to 0·75)	1·63 (0·41 to 8·13)	0·24 (0·095 to 0·43)	0·95 (0·88 to 1·05)
LASSO	Restricted	0·66 (0·63 to 0·69)	1·19 (0·43 to 4·39)	0·11 (–0·06 to 0·33)	0·84 (0·67 to 1·02)
LASSO	Extensive	0·68 (0·65 to 0·71)	1·26 (0·42 to 5·50)	0·05 (–0·09 to 0·24)	0·78 (0·65 to 0·93)

Discrimination and continuous QoL score prediction performances were assessed with bootstrap .632+ estimator and calibration estimates were obtained with the .632 estimator. Models were assessed with two different sets of predictors (restricted and extensive). CIL=calibration-in-the-large. LASSO=least absolute shrinkage and selection operator. MAE=mean absolute error. QoL=quality of life.

Table 2: **Predictive performances**

probability of QoL decline or death (joint model), (C) conditional QoL decline, and (D) the probability of being deceased at 36 months after diagnosis. The models present skewed prediction distributions, in agreement with the distributions observed in the development data. Moderate miscalibration occurred at continuous outcome extremes and in the conditional QoL decline model with systematic risk overestimation. Miscalibriation was defined as a noticiable deviation from the diagonal of the calibration curve. Weak miscalibration was observed for the joint model with risk underestimation probabilities ranging from 0·3 to 0·4. The survival model showed good calibration. Overall, the models were reasonably well calibrated with good alignment with the observed data.

In external validation, LASSO and XGboost with the restricted set of parameters models performed similarly to their internal validation performance for the continuous QoL with MAEs of 12·97 and 12·28, and R^2 of 0·353 and 0·414, respectively. Discrimination was higher than the discrimination observed in internal validation with C-statistics of 0·799 and 0·795 for LASSO and XGBoost respectively. The models presented overall good calibration in the external data but with large confidence intervals, with slight indication of systematic risk of underestimation for LASSO and overestimation for XGBoost (figure 3). The survival submodel was validated with an area under the curve of 0·60, and calibration metrics including an observed-to-expected ratio of 0·64, calibration-in-the-large of 0·15, and calibration slope of 0·64. The calibration curve (appendix p 21) showed an underprediction of mortality, particularly close to the mean predicted mortality (approximately 0·1).

SHAP values quantified the effect of the covariates on the prediction for each patient. Figure 4 presents the SHAP values for all individuals in the internal dataset for (A) mortality, conditional QoL with (B) LASSO, (C) XGBoost, and (D) the SHAP interactions from XGBoost. Common predictors of mortality, such as age and TNM stage, are associated with increased probability of death. Conditional QoL is mainly linked with QoL measured at earlier timepoints, such as at baseline, suggesting strong autocorrelation. The strongest interactions were observed between QoL at 12 months after diagnosis and other functional scales, household income, BMI, alcohol consumption, and age.

Discussion

This study introduces a joint probability approach for prognostic models for conditional outcomes that can be tailored to answer relevant probabilistic questions at inference time. We applied this framework and developed a prognostic model for QoL in a large cohort of patients with head and neck cancer, which provided robust datasets for model development and validation. The results indicate that the joint model, which combines predictions of survival and QoL, offers a more comprehensive prognostic picture compared with models that only consider the conditional outcome. External validation showed good discriminative performance. However, future research should focus on external validation of the model in different populations (eg, with more ethnic diversity) and settings (eg, countries with other health-care systems) to confirm transportability. More generally, we have formally detailed using probability theory why the handling of post-baseline events such as death cannot be decoupled from the prediction target of interest.

We developed and made available the BD4QoLPredict API and dashboard to facilitate access to the models either programmatically or via a user-friendly interface.

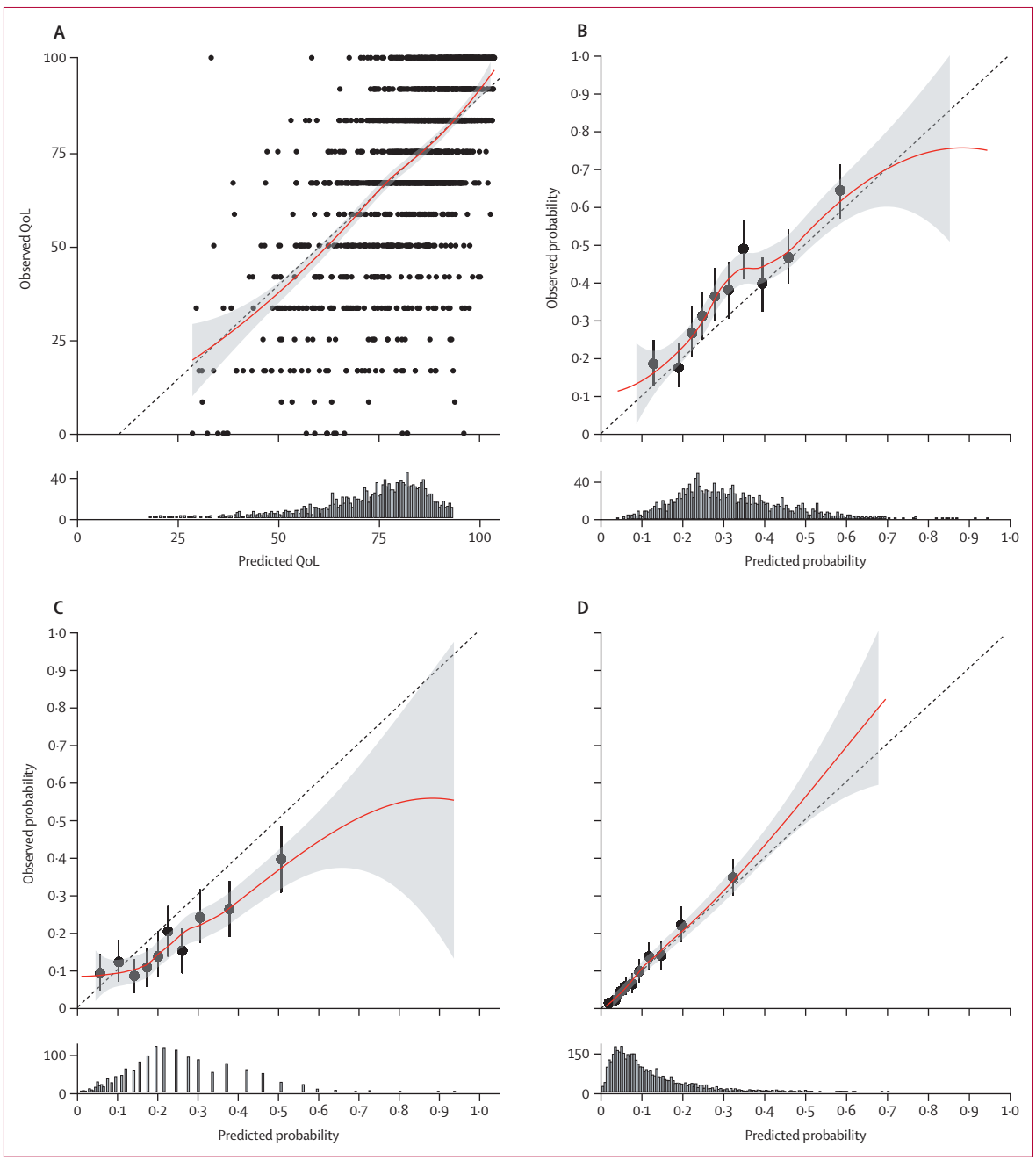

Figure 2: **Calibration plots for the predictive models at 3 years after diagnosis**
(A) Continuous outcome QoL (measured on a scale of 0 to 100). (B) Probability of decline or death. (C) Conditional probability of QoL decline. (D) Probability of survival. In each plot, the x axis represents the predicted values, whereas the y axis represents the observed values. The black dots in (B), (C), and (D) represent the estimated observed probabilities for different bins of predicted probabilities and 95% CIs. The solid red lines are smoothed curves obtained with LOESS regression, and the grey shaded areas denote the 95% CIs. The 45° diagonal lines are the reference for perfect calibration. At the bottom of each calibration the predicted QoL and the probability histograms are shown (the y axis shows frequency). LOESS=locally estimated scatterplot smoothing. QoL=quality of life.

BD4QoLPredict sets a high standard as one of the first ready-to-use prognostic tools developed following rigorous methodology to minimise bias and to provide probabilistic QoL predictions for patients with head and neck cancer. Many previous studies were aimed at prognostic factor discovery in this population,[22,29] but no previous prognostic tool is available.

The joint probability model has considerable potential for clinical use. The main contribution lies in identifying prognostic factors that are relevant and readily available in routine practice (eg, baseline and early post-treatment QoL scores, tumour stage, comorbidities, and lifestyle factors) that most strongly influence long-term outcomes, thereby supporting early risk stratification and targeted

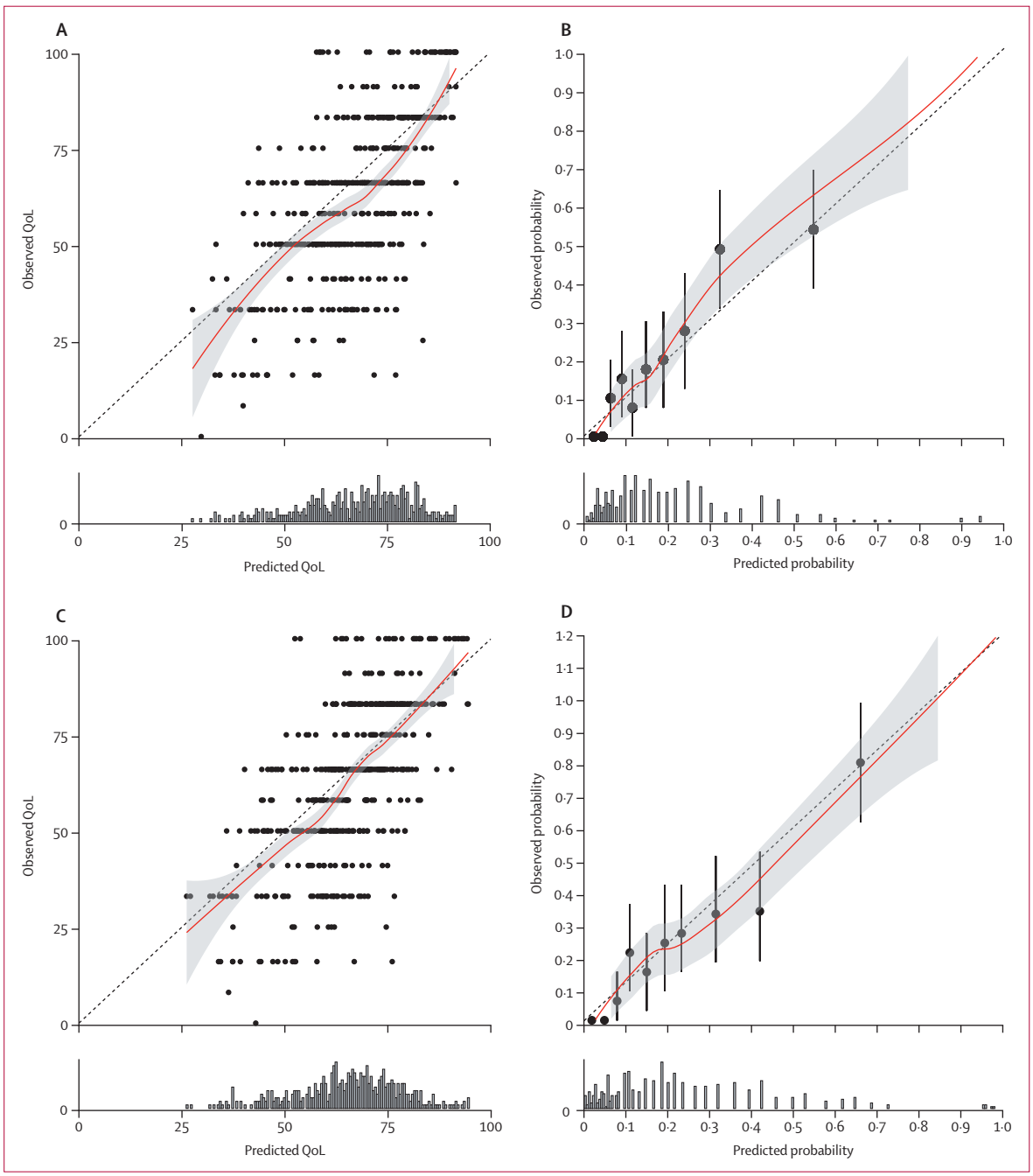

Figure 3: **External validation for models developed with the limited set of predictors**
Panels (A) and (B) depict calibration curves for the continuous QoL and conditional QoL decline for LASSO, whereas (C) and (D) show calibration for XGBoost. The black dots in (C) and (D) represent the estimated observed probabilities for different bins of predicted probabilities and 95% CIs. The solid red lines are smoothed curves obtained with LOESS, and the grey shaded areas denote the 95% CIs. The 45° diagonal lines are the reference for perfect calibration. At the bottom of each calibration the predicted QoL and the probability histograms are shown (the y axis shows frequency). LASSO=least absolute shrinkage and selection operator. LOESS=locally estimated scatterplot smoothing. QoL=quality of life.

survivorship interventions. By providing predictions for both survival and QoL decline together, the model could help clinicians identify patients at high risk of adverse outcomes who could benefit from early interventions aimed at improving both survival and quality of life. This approach could also inform health-care resource planning, ensuring that efforts are directed towards patients with the greatest need. For example, the model could support early referral to supportive and rehabilitative services (eg, swallowing or speech therapy, nutritional support, and

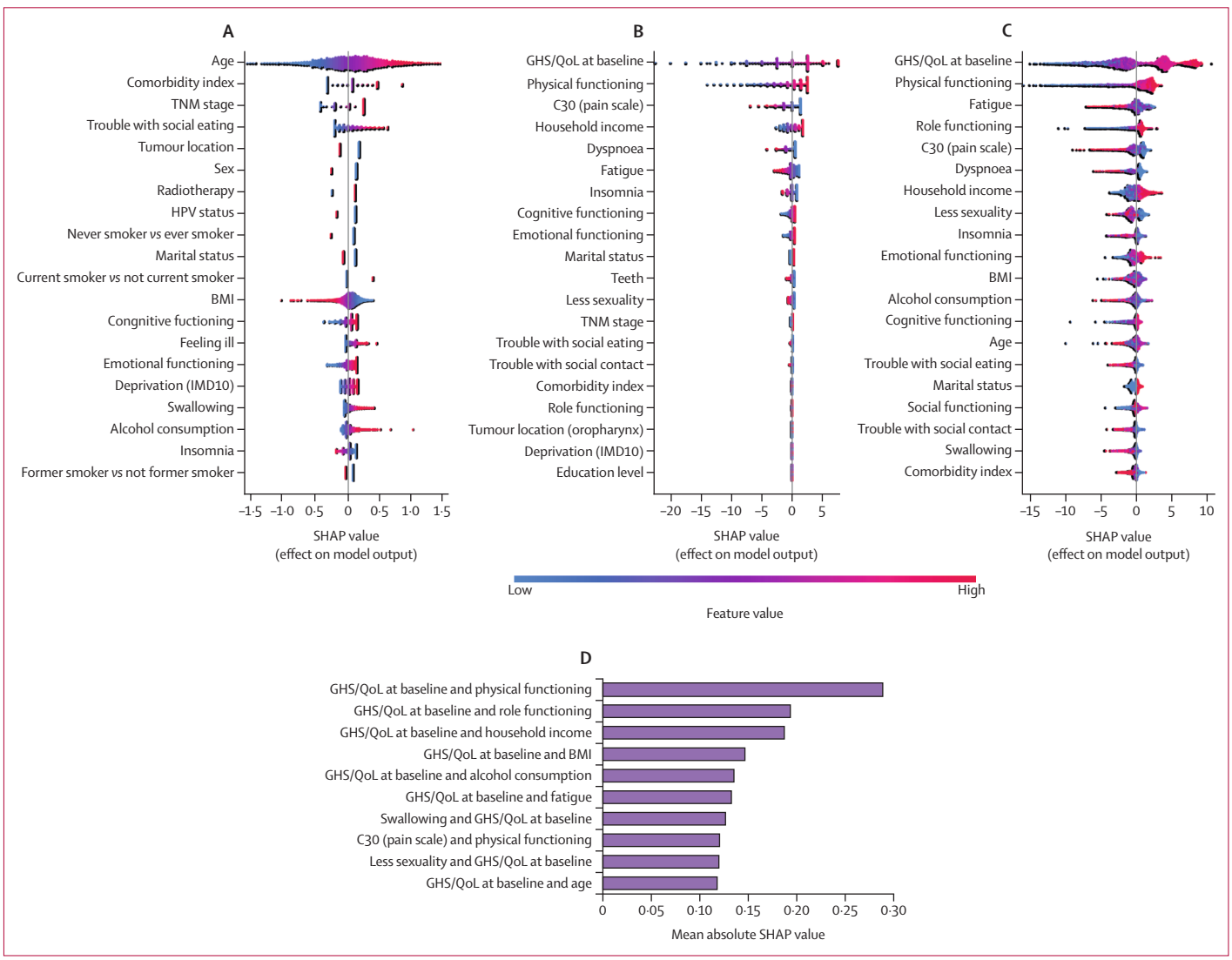

Figure 4: **SHAP values associated with prediction models for mortality (A), conditional QoL with LASSO (B), XGBoost (C), and the respective top 1% of mean absolute SHAP interaction values (D)**
GHS=Global Health Status. HPV=human papillomavirus. IMD10=Index of Multiple Deprivation 2010. LASSO=least absolute shrinkage and selection operator. QoL=quality of life. SHAP=Shapley Additive Explanations.

psychosocial interventions) and enable closer follow-up for patients identified as being at high risk. Even if the underlying cause cannot be reversed, such proactive management can mitigate symptom burden and improve patients' perceived wellbeing.[30] Moreover, although the current tool predicts global QoL, future developments might extend to domain-specific outcomes such as physical functioning and social functioning, thus offering more targeted guidance for clinical interventions.

A key strength of the joint modelling approach for conditional outcomes is that a probabilistic interpretation of the predictions is maintained. By contrast, commonly used alternatives—such as restricting model development to individuals in whom the conditional outcome is observed, or imputation of the conditional outcome for those in whom the conditional outcome is not defined[31]—do not retain this property. Restricting model development to individuals in whom the conditional outcome is observed results in conditional probabilities or classifications, but this might not be explicitly acknowledged and could thus result in invalid interpretations. The imputation approach makes the interpretation of the predictions difficult because values are imputed for individuals for whom the conditional outcome is not defined.

Another key strength of the joint modelling approach is that it results in a flexible model that, when coupled with an API and dashboard, can be tailored to the relevant questions in the field. One can ask questions conditional on the post-baseline event occurring or about the joint probability of the post-baseline event occurring and about some state or range for the conditional outcome.

A limitation of the joint modelling approach could be in datasets in which the relevant post-baseline event is not observed and in which it is impossible to distinguish between outcomes that are missing due to being conditionally defined or missing due to other causes.

An important feature of the joint modelling approach is that it is probabilistic, which provides direct interpretation for the predictions compared with methods that provide only score-based binary classification. The model can predict the continuous measurement with prediction intervals, the probability of the outcome lying in a post-defined range, and the probability of the outcome not being observed.

One key strength in our application is the use of a large, well characterised cohort of patients with head and neck cancer. By developing multivariable models in a heterogeneous population of patients with head and neck cancer, the model can provide personalised predictions according to the patient's characteristics. The study also used advanced statistical techniques, including conformal prediction and SHAP, to ensure robust model development and interpretation. We followed high methodological standards for model development and performance assessment with internal and external validation, by estimating important metrics such as calibration and using bias-corrected bootstrap estimators.

However, there are some limitations to consider. The sample size, although substantial, could still pose a risk of overfitting, particularly for the extended set of predictors. Currently, the model is developed for QoL using the EORTC instrument, which is comprehensive and commonly used in European clinical trials. However, future work could include training the model on data using another QoL instrument (eg, the EQ-5D-5L). We would also suggest monitoring model performance over time, especially as improvements are made such as earlier diagnosis and other treatment options such as immunotherapy or app-based QoL-monitoring tools emerge.[21]

The study also faced challenges with missing data, but these were addressed using imputation methods in the model development and external validation, and propensity scores for model evaluation, which indicated consistent model performance in patients similar to and dissimilar to patients with missing outcomes. These methods mitigate concerns about model validity that otherwise would remain due to a high degree of missingness of QoL at the time of prediction. In addition, future research can explore different optimisation strategies for the joint model estimation by combining the objective functions of the submodels rather than estimating them separately.

Finally, the model performance observed was modest. However, this aligns with the broader challenge of the prediction of clinical and biological outcomes, which are part of complex systems, and suggests that although the model has predictive capacity, it should be used as a supplementary tool alongside clinical judgement. The model's ability to predict both survival and QoL provides a more holistic view of patient prognosis, unlike a conditional approach that assumes survival. This comprehensive perspective is crucial for clinical decision making and resource allocation.

In conclusion, this study presents a novel joint probability approach to prediction modelling for conditional outcomes, with application in patients with head and neck cancer. We present a comprehensive prognostic tool that could aid in clinical decision making and resource allocation. We underscore the importance of recognising conditional outcomes when developing prognostic prediction models.

Contributors

ML, AF, MM-S, and LL contributed to the conceptualisation. ML, AF, MM-S, AB-J, AA, and EIFF contributed to the methodology. ML and AF contributed to the supervision. ML and MM-S drafted the first draft of the manuscript. SS, LL, ST, and MP contributed data. MM-S and EIFF contributed to the data analysis. ML, MM-S, EIFF, AB-J, AA, LL-P, and IA had full access to and verified all study data. Access for all authors to the raw data was not feasible due to legal restrictions. SS and KT provided expertise on quality-of-life measurement. SC, LL, ST, and MP provided clinical expertise on patients with head and neck cancer. LL-P, IA, AB-J, MFC-U, G-F, and AA provided expertise on the evaluation of machine learning models. All authors had access to the data presented in this study. All authors reviewed and approved the manuscript and had final responsibility for the decision to submit for publication.

Declaration of interests

SS reports receiving honoraria for reviewing journal papers for the Lilly Quality of Life Prize, outside of this work. LL declares research funds to their institute for clinical studies from AstraZeneca, BMS, Boehringer Ingelheim, Celgene International, Eisai, Exelixis, Debiopharm International, Hoffmann-La Roche, IRX Therapeutics, Medpace, Merck-Serono, MSD, Novartis, Pfizer, Roche, Buran, and Alentis, and occasional fees for participation as a speaker at conferences or congresses or as a scientific consultant for advisory boards from Alentis, AstraZeneca, Bayer, MSD, Merck-Serono, AccMed, and Neutron Therapeutics. SC declares occasional fees for participation as a speaker at conferences or congresses from AccMed; and support for attending meetings or travel from AccMed, MultiMed Engineers, and Care Insight. ML declares being on the data safety monitoring board for several investigator-initiated studies at Oslo University Hospital. KJT declares honoraria and travel expenses from the European Organisation for Research and Treatment of Cancer. All other authors declare no competing interests.

Data sharing

Non-sensitive data generated during model development are shared in this manuscript and the appendix. All datasets used in this study are hosted by the Services for Sensitive Data at the University of Oslo (Oslo, Norway). Access to the data could be granted by the data owners upon application. The data that support the findings of this study are available from Head and Neck 5000 (HN5000), with further information on the HN5000 website (https://www.headand neck5000.org.uk/information-for-researchers). For information on access to the external validation cohorts, contact LL and SS.

Acknowledgments

The BD4QoL project (LL was the principal investigator and funding was given to all co-authors), in the frame of which this work is being conducted, has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement number 875192. MM-S received funding from the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions Grant, agreement number 801133 (Scientia fellowship). The Italian validation data used in this study are from the BD2Decide project (LL was the principal investigator), which was funded by the EU's Horizon 2020 research and innovation programme under grant number 689715.

This publication presents data from the HN5000 study. The study was a component of independent research funded by the National Institute for Health and Care Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-070-10034). The views expressed in this publication are those of the authors and are not necessarily those of the UK National Health Service (NHS), the NIHR, or the Department of Health and Social Care. Core funding was also provided through awards from Above & Beyond, a component of the University Hospitals Bristol and Weston Research Capability Funding. The 3-year follow-up period was supported by a Cancer Research UK Programme Grant, the Integrative Cancer Epidemiology Programme (grant number C18281/A19169).

The German validation datasets (SS was the principal investigator) were funded by the German Federal Ministry of Education and Science (grant number 7DZAIQTX), the German Cancer Aid (grant numbers #106654, #107440, #108758, and #109604), and the the European Organisation for Research and Treatment of Cancer (grant number 001/2014).

The data are stored in compliance with the General Data Protection Regulation legislation in the secure server for sensitive data at the University of Oslo (TSD/USIT) and access is granted to authorised collaborators included in the ethical approval.

This work was performed on the Services for Sensitive Data (TSD) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo IT Department. The statistical analysis and data harmonisation were performed on resources provided by Sigma2, the national infrastructure for high performance computing and data storage in Norway. The authors thank the HN5000 executive group and study pathologists and the researchers and clinicians who designed the HN5000 study; the research, laboratory, and clinical staff who supported the conduct of the study; and the people with head and neck cancer who took part in the study. During the preparation of this work the authors used GTP-UiO with GPT-4o in order to reduce the word count in a few sentences. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1 Pinto M, Marotta N, Caracò C, Simeone E, Ammendolia A, de Sire A. Quality of life predictors in patients with melanoma: a machine learning approach. *Front Oncol* 2022; **12:** 843611.

2 Colantuoni E, Li X, Hashem MD, Girard TD, Scharfstein DO, Needham DM. A structured methodology review showed analyses of functional outcomes are frequently limited to "survivors only" in trials enrolling patients at high risk of death. *J Clin Epidemiol* 2021; **137:** 126–32.

3 Granholm A, Anthon CT, Kjær MN, et al. Patient-important outcomes other than mortality in contemporary ICU trials: a scoping review. *Crit Care Med* 2022; **50:** e759–71.

4 Oeyen S, Vermeulen K, Benoit D, Annemans L, Decruyenaere J. Development of a prediction model for long-term quality of life in critically ill patients. *J Crit Care* 2018; **43:** 133–38.

5 Révész D, van Kuijk SMJ, Mols F, et al. External validation and updating of prediction models for estimating the 1-year risk of low health-related quality of life in colorectal cancer survivors. *J Clin Epidemiol* 2022; **152:** 127–39.

6 Sammut L, Ward M, Patel N. Physical activity and quality of life in head and neck cancer survivors: a literature review. *Int J Sports Med* 2014; **35:** 794–99.

7 Crowder SL, Douglas KG, Yanina Pepino M, Sarma KP, Arthur AE. Nutrition impact symptoms and associated outcomes in post-chemoradiotherapy head and neck cancer survivors: a systematic review. *J Cancer Surviv* 2018; **12:** 479–94.

8 Glare PA, Davies PS, Finlay E, et al. Pain in cancer survivors. *J Clin Oncol* 2014; **32:** 1739–47.

9 Bossi P, Giusti R, Tarsitano A, et al. The point of pain in head and neck cancer. *Crit Rev Oncol Hematol* 2019; **138:** 51–59.

10 Yu J, Smith J, Marwah R, Edkins O. Return to work in patients with head and neck cancer: systematic review and meta-analysis. *Head Neck* 2022; **44:** 2904–24.

11 Klein J, Livergant J, Ringash J. Health related quality of life in head and neck cancer treated with radiation therapy with or without chemotherapy: a systematic review. *Oral Oncol* 2014; **50:** 254–62.

12 Taylor K, Singer S. Long-term quality of life in head and neck cancer patients: a systematic review. *Die Onkologie* 2019; **25:** 125–31.

13 Taylor K, Krüger M, Singer S. Long-term toxicity among head and neck cancer patients—a systematic review. *Die Onkologie* 2021; **27:** 145–49.

14 Taylor KJ, Amdal CD, Bjordal K, et al. Serious long-term effects of head and neck cancer from the survivors' point of view. *Healthcare (Basel)* 2023; **11:** 906.

15 Taquet V, Blot V, Morzadec T, Lacombe L, Brunel N. MAPIE: an open-source library for distribution-free uncertainty quantification. 2022. http://arxiv.org/abs/2207.12274 (accessed Sept 16, 2024).

16 Ness AR, Waylen A, Hurley K, et al, and the Head and Neck 5000 Study Team. Establishing a large prospective clinical cohort in people with head and neck cancer as a biomedical resource: head and neck 5000. *BMC Cancer* 2014; **14:** 973.

17 Ness AR, Waylen A, Hurley K, et al, and the Head and Neck 5000 Study Team. Recruitment, response rates and characteristics of 5511 people enrolled in a prospective clinical cohort study: head and neck 5000. *Clin Otolaryngol* 2016; **41:** 804–09.

18 Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993; **85:** 365–76.

19 Bjordal K, Hammerlid E, Ahlner-Elmqvist M, et al. Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. *J Clin Oncol* 1999; **17:** 1008–19.

20 Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Life Group. EORTC QLQ-C30 Scoring Manual, 3rd edn. Brussels: European Organisation for Research and Treatment of Cancer, 2001.

21 Cavalieri S, Vener C, LeBlanc M, et al, and the BD4QoL Consortium. A multicenter randomized trial for quality of life evaluation by non-invasive intelligent tools during post-curative treatment follow-up for head and neck cancer: clinical study protocol. *Front Oncol* 2023; **13:** 1048593.

22 Alonso I, Lopez-Perez L, Guirado JCM, Cabrera-Umpierrez MF, Arredondo MT, Fico G. Data analytics for predicting quality of life changes in head and neck cancer survivors: a scoping review. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1–5, 2021.

23 Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; **368:** m441.

24 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 2015; **102:** 148–58.

25 Cordier T, Blot V, Lacombe L, Morzadec T, Capitaine A, Brunel N. Flexible and systematic uncertainty estimation with conformal prediction via the MAPIE library. In: Papadopoulos H, Nguyen KA, Boström H, Carlsson L, eds. Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications. PMLR, 2023: 549–81.

26 Vovk V, Shen J, Manokhin V, Xie M-ge. Nonparametric predictive distributions based on conformal prediction. In: Gammerman A, Vovk V, Luo Z, Papadopoulos H, eds. Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications. PMLR, 2017: 82–102.

27 Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *J Am Stat Assoc* 1997; **92:** 548.

28 Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Stat* 1983; **37:** 36–48.

29 Moubayed SP, Sampalis JS, Ayad T, et al. Predicting depression and quality of life among long-term head and neck cancer survivors. *Otolaryngol Head Neck Surg* 2015; **152:** 91–97.

30 Temel JS, Greer JA, Muzikansky A, et al. Early palliative care for patients with metastatic non-small-cell lung cancer. *N Engl J Med* 2010; **363:** 733–42.

31 Nijman S, Leeuwenberg AM, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J Clin Epidemiol* 2022; **142:** 218–29.