ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Preference-Based Assessments

Gradient Boosted Tree Approaches for Mapping European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 Onto 5-Level Version of EQ-5D Index for Patients With Cancer

Yasuhiro Hagiwara, PhD, MPH, Takeru Shiroiwa, PhD, MPH, Naruto Taira, MD, PhD, Takuya Kawahara, PhD, MPH, Keiko Konomura, PhD, Shinichi Noto, PhD, Takashi Fukuda, PhD, Kojiro Shimozuma, MD, PhD

Objectives: This study aimed to develop direct and response mapping algorithms from the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 onto the 5-level version of EQ-5D index based on the gradient boosted tree (GBT), a promising modern machine learning method.

Methods: We used the Quality of Life Mapping Algorithm for Cancer study data (903 observations from 903 patients) for training GBTs and testing their predictive performance. In the Quality of Life Mapping Algorithm for Cancer study, patients with advanced solid tumor were enrolled, and the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 and 5-level version of EQ-5D were simultaneously evaluated. The Japanese value set was used for direct mapping, whereas the Japanese and US value sets were used for response mapping. We trained the GBTs in the training data set (80%) with cross-validation and tested the predictive performance measured by the root mean squared error (RMSE), mean absolute error (MAE), and mean error in the test data set (20%).

Results: The RMSE and MAE in the test data set were larger in the GBT approaches than in the previously developed regression-based approaches. The mean error in the test data set tended to be smaller in the GBT approaches than in the previously developed regression-based approaches.

Conclusions: The predictive performances in the RMSE and MAE did not improve by the GBT approaches compared with regression approaches. The flexibility of the GBT approaches had the potential to reduce overprediction and underprediction in poor and good health, respectively. Further research is needed to establish the role of machine learning methods in mapping a nonpreference-based measure onto health utility.

Keywords: 5-level version of EQ-5D, European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30, gradient boosted tree, health utility, machine learning, mapping.

VALUE HEALTH. 2022; ■(■):■–■

Introduction

Mapping from a nonpreference-based measure onto health utility is a valuable tool in health economic evaluation when a nonpreference-based measure is used instead of directly measuring health utility.[1-3] A mapping algorithm predicts health utility from responses to a nonpreference-based measure and enables analysts to estimate quality-adjusted life-years, an important effectiveness measure in health economic evaluation. Mapping algorithms from a nonpreference-based measure onto health utility can be classified into 2 types; a direct mapping algorithm directly predicts health utility, whereas a response mapping algorithm predicts responses to a preference-based measure and indirectly estimates health utility from predicted responses. Both types of mapping algorithms are usually based on classical regression models including linear regression models, tobit models, and generalized linear models for direct mapping and ordinal logistic and ordinal probit models for response mapping.[3]

Many mapping algorithms incur overprediction in poor health and underprediction in good health.[1,4,5] This problem stems from the inflexibility of regression models that does not account for possibly complex nature of association between health utility and a nonpreference-based measure, such as a nonlinear relation between health utility and subscales in a nonpreference-based measure as well as interaction between subscales in a nonpreference-based measure. The inflexibility of regression models that have been used for mapping a nonpreference-based measure onto health utility may also result in large prediction errors, as measured by the root mean squared error (RMSE) and mean absolute error (MAE).

Modern machine learning methods can overcome these problems in conventional mapping algorithms based on inflexible regression modeling. Machine learning provides flexible methods to approximate a function[6] and hence has the potential to provide more accurate prediction of health utility than conventional regression approaches. Although modern machine learning methods have led to improvements in many fields of medicine,[7]

they are rarely applied to mapping from a nonpreference-based measure onto health utility.[8]

As a promising machine learning method for mapping a nonpreference-based measure onto health utility, we focused on a gradient boosted tree (GBT), also known as gradient boosting tree or gradient tree boosting.[9,10] GBT uses boosting as an ensemble learning method and applies it to decision tree. The basic idea of ensemble learning is to combine multiple base models to obtain better prediction. In a GBT algorithm, decision trees are sequentially fitted to residuals (ie, differences between observed and predicted outcomes; more generally, the negative gradient of a loss function) and updates the prediction and residuals. Decision trees can automatically incorporate variable selection, nonlinear relations between outcome and predictor variables, and interaction between predictor variables, whereas boosting improves prediction in areas where it does not perform well while using some techniques to reduce overfitting. As a result of its good performance, GBT has become popular in medical research (see Appendix in Supplemental Material found at https://doi.org/10.1016/j.jval.2022.07.020 for results of the literature search). In a review article on machine learning in health outcomes research,[11] the authors state that GBT "often achieves the highest performance (lowest error between predicted and observed outcomes) among modern machine learning methods for tabular data." Thus, GBT is a promising approach for mapping a nonpreference-based measure onto health utility. Nevertheless, to the best of our knowledge, GBT has not been applied to develop mapping algorithms from a nonpreference-based measure onto health utility.

This study aimed to develop direct and response mapping algorithms using GBTs and compare their predictive performances with those of previously developed mapping algorithms based on regression approaches. We focused on the oncology area and developed mapping algorithms from the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ-C30), a common cancer-specific health-related quality of life questionnaire, onto the 5-level version of EQ-5D (EQ-5D-5L) index.

Methods

Study Data

We used data from the Quality of Life Mapping Algorithm for Cancer (QOL-MAC) study to train and test GBT-based direct and response mapping algorithms. Details of the QOL-MAC study has been reported elsewhere.[12] In brief, the QOL-MAC study was a cross-sectional study conducted to develop mapping algorithms for the EORTC QLQ-C30 and the Functional Assessment of Cancer Therapy General onto the EQ-5D-5L index. A total of 1031 patients who received drug therapy for locally advanced, metastatic, or recurrent solid tumor were enrolled from 14 hospitals in Japan between 2018 and 2019. Other inclusion criteria were age (≥ 20 years) and with an Eastern Cooperative Oncology Group performance status of 0 to 3. The study protocol of the QOL-MAC study was approved by each participating hospital. All enrolled patients provided a written informed consent before their enrollment in the study.

Instruments

In the QOL-MAC study, the EQ-5D-5L was assessed using the Japanese version of the EQ-5D-5L questionnaire.[13,14] The EQ-5D-5L includes 5 items: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. It requires patients to assign a 5-point status level to each of the 5 items: no problems, slight problems, moderate problems, severe problems, or extreme problems (the wording is slightly different among the 5 items). We used the value set of the Japanese population for direct mapping[14] because it was used to develop regression-based direct mapping algorithms using the QOL-MAC study data.[12] For indirect mapping, we used the value set for the Japanese population and that for the population of the United States as a Western country's value set.[15]

The Japanese version of the EORTC QLQ-C30 questionnaire (version 3) was used in the QOL-MAC study.[16] The EORTC QLQ-C30 has 30 items and the responses to these items were converted into 5 functioning subscale scores (physical, role, emotional, cognitive, and social), 3 symptom subscales calculated using multiple items (fatigue, nausea, and vomiting, and pain), 6 symptom subscales using a single item (dyspnea, insomnia, appetite loss, constipation, diarrhea, and financial difficulties), and a global health status score. Higher scores on the functioning subscales and global health status indicate better health conditions, whereas higher scores on symptom subscales indicate more severe symptoms. All scores range from 0 to 100.

Preliminary Analysis

We used the analysis population for the EORTC QLQ-C30 in the QOL-MAC study,[12] which was defined as eligible patients having both the EQ-5D-5L index and all 15 subscale scores of the EORTC QLQ-C30 (903 patients of 1031 enrolled patients). We randomly divided this analysis population into the training data set (722 patients, 80%), which was used to develop GBT-based direct and response mapping algorithms, and the test data set (181 patients, 20%), which was used to evaluate the predictive performance of the developed mapping algorithms.

Demographic characteristics, clinical characteristics, and distributions of the EQ-5D-5L and EORTC QLQ-C30 of patients in the training and test data sets were summarized. To evaluate the conceptual overlap between the EQ-5D-5L and EORTC QLQ-C30, we calculated Spearman's rank correlation coefficients between the EQ-5D-5L items and EORTC QLQ-C30 subscale scores. Data preparation and preliminary analyses were conducted using Base SAS and SAS/STAT version 9.4 software of the SAS System for Windows (SAS Institute, Cary, NC, USA).

Training GBTs

We developed GBT-based direct and response mapping algorithms using the training data set. A GBT is a type of ensemble learning method called boosting that uses a decision tree as a base learner.[9-11] A GBT sequentially fits decision trees to residuals (ie, differences between observed and predicted outcomes; more generally, the negative gradient of a loss function) and updates the prediction. Because a decision tree is used as a base learner, it can automatically incorporate variable selection, nonlinearity, and interaction.

For both the direct and response mapping algorithms, all 15 subscale scores in the EORTC QLQ-C30, age, and sex were used as predictor variables. To develop a direct mapping algorithm, we used the EQ-5D-5L index based on the value set of the Japanese population as an outcome and used the RMSE as a loss function to be minimized in training. When developing the response mapping algorithm, we used responses to the 5 items of the EQ-5D-5L as outcomes and the negative multinomial log likelihood as loss functions. In addition to the responses to the 5 items of the EQ-5D-5L, we predicted not full health for response mapping using negative binary log likelihood as a loss function, because the Japanese value set had a disutility term for being not full health in addition to the 5 items.

GBT has several hyperparameters that must be tuned to obtain good predictive performance. First, we set the learning rate to 0.01. In GBT, prediction by decision tree in each boosting iteration was multiplied by the learning rate ranging from 0 to 1. A lower learning rate can prevent overfitting, but requires a larger number of boosting iterations. Learning rates of < 0.1,[10] 0.01, and 0.001[17] are recommended in machine learning textbooks. Then, we conducted a grid search of the optimal pair of hyperparameters of the maximum depth of the tree in each boosting iteration, row subsampling proportion, and column subsampling proportion. The maximum depth of the tree corresponded to the maximum interaction order. For example, a tree with a depth of 1 is a stump and no interaction between predictor variables is incorporated in the prediction, whereas a tree with a depth of 3 can have 8 leaves, and the interaction among 3 predictor variables can be incorporated. Because the interaction order of 3 to 7 works well empirically in many problems,[10] we searched the best maximum depth of the tree from 1 to 7 by 1. In row and column subsampling, observations and predictor variables are randomly sampled without replacement at each boosting iteration, respectively. These processes can reduce the correlation among predictions in boosting iterations and hence produce accurate predictions as a whole.[18] We searched the best row and column sampling proportion from 0.25 to 1 by 0.25.

The grid search to tune the hyperparameters was conducted using ninefold cross-validation in the training data set. In cross-validation, a total of $7 \times 4 \times 4 = 112$ combinations of hyperparameters were evaluated. We ran a maximum of 10 000 boosting iterations with early stopping after 100 consecutive boosting iterations did not improve a cross-validated loss function. After cross-validation found the optimal combination of hyperparameters, we trained the final GBTs with the optimal values of hyperparameters using the whole training data set and obtained prediction models for the EQ-5D-5L index (for direct mapping) and probabilities of the 5 levels of the 5 items in the EQ-5D-5L and not full health (for response mapping). Because GBT does not provide regression coefficients for predictor variables, we calculated predictor variable importance measures called the gain, frequency, and cover in GBT analysis for interpretation of predictive performances.

Evaluating Predictive Performances

We calculated predictive performance measures in the training and test data sets for mapping algorithms that were newly developed using GBTs here and those previously developed using regression approaches from the QOL-MAC study data. The recommended regression-based direct mapping algorithm was based on the 2-part beta regression,[12] which predicted full health by logistic regression and modeled the remaining EQ-5D-5L index using beta regression. The recommended regression-based response mapping algorithm was based on ordinal logistic regression,[12] which predicted probabilities of the 5 levels of the 5 items in the EQ-5D-5L by ordinal logistic regression, and the EQ-5D-5L index was calculated using these probabilities. The detailed methods for obtaining the predicted EQ-5D-5L index from these mapping algorithms were reported.[12] We did not consider regression-based mapping algorithms that were developed outside of the QOL-MAC study to compare GBT approaches with regression approaches without being affected by differences in the data sets that were used to develop mapping algorithms. The performance of the direct mapping algorithms was evaluated for the Japanese value set, whereas those of response mapping algorithms were evaluated for both the Japanese and US value sets. We evaluated the predictive performance based on the RMSE, MAE, and mean error (ME) (see Appendix in Supplemental Material found at https://doi.org/10.1016/j.jval.2022.07.020 for detailed calculation of these performance measures). The RMSE and MAE were used to evaluate the average distance between the observed and predicted EQ-5D-5L index, whereas ME was used to evaluate overestimation and underestimation of the EQ-5D-5L index. These measures were calculated for the whole training and test data sets and subgroups defined by the global health status score, which reflects the general health condition of the patients.

GBT analyses and performance evaluations were conducted using R 4.0.5 (R Core Team, Vienna, Austria). GBTs were fitted using the R package xgboost.[19,20]

Results

Patients and Descriptive Data

A total of 1031 patients were enrolled in the QOL-MAC study. Of the 1029 eligible patients, 903 patients had the EQ-5D-5L index and all subscale scores of the EORTC QLQ-C30. The training data set included 722 patients, whereas the test data set included 181 patients (Table 1). The median age was 67 years in the training data set and 68 years in the test data set. The most frequent cancer type was lung cancer, followed by colorectal cancer in the 2 data sets. Patients with an Eastern Cooperative Oncology Group PS of 2 or higher amounted to 9.4% in the training data set and 8.3% in the test data set.

The distributions of the EQ-5D-5L index and subscale scores of the EORTC QLQ-C30 in the training and test data sets are presented in Table 2 (histograms were provided as Appendix Figs. 1-3 in the Supplemental Material found at https://doi.org/10.1016/j.jval.2022.07.020). We provide the data on the response to each EQ-5D-5L item in Appendix Table 1 in the Supplemental Material found at https://doi.org/10.1016/j.jval.2022.07.020. The distributions of the EQ-5D-5L index and subscale scores were generally matched between the training and test data sets. The rank correlations between subscale scores of the EORTC QLQ-C30 and the responses to the 5 items in the EQ-5D-5L are reported in Appendix Table 2 in the Supplemental Material found at https://doi.org/10.1016/j.jval.2022.07.020. The rank correlations between each item in the EQ-5D-5L and subscale in EORTC QLQ-C30 were qualitatively similar between the 2 data sets. All items had at least one subscale score with an absolute correlation larger than 0.4.

Developing GBT-Based Mapping Algorithms

After the ninefold cross-validation, the optimal values of the hyperparameters were chosen. The optimal tree depth was 2 for index and 1 for mobility, self-care, usual activities, pain/discomfort, anxiety/depression, and not full health (see Appendix Table 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.07.020 for the optimal values of row and column subsampling proportions). Figure 1 shows the loss functions related to GBTs for direct and response mapping in the cross-validation. Loss functions in the training set consistently decreased over boosting iterations, whereas those in the validation set hit the bottom, providing the optimal number of boosting iterations: 467 for index, 782 for mobility, 628 for self-care, 1136 for usual activities, 737 for pain/discomfort, 1013 for anxiety/depression, and 621 for not full health.

Figure 2 shows the predictor variable importance measured by the gain for direct and response mapping in the training data set with the optimal hyperparameter values. The most important variable for direct mapping was physical functioning, followed by role functioning. The most important variables for response

Table 1. Patient characteristics in the training data set and the test data set.

Characteristic	Training data set (n = 722)	Test data set (n = 181)
Age (years)	67 (58-74)	68 (59-74)
Sex		
Male	395 (54.7)	94 (51.9)
Female	327 (45.3)	87 (48.1)
Hospitalization		
Yes	152 (21.1)	34 (18.8)
No	570 (78.9)	147 (81.2)
ECOG performance status		
0	348 (48.2)	103 (56.9)
1	306 (42.4)	62 (34.3)
2	52 (7.2)	11 (6.1)
3	16 (2.2)	4 (2.2)
Unknown (0, 1, 2, or 3)	0 (0)	1 (0.6)
Tumor type		
Lung cancer	247 (34.2)	70 (38.7)
Stomach cancer	51 (7.1)	14 (7.7)
Colorectal cancer	182 (25.2)	40 (22.1)
Breast cancer	92 (12.9)	20 (11.0)
Other solid tumors	149 (20.6)	37 (20.4)
Stage at diagnosis		
I	39 (5.4)	10 (5.5)
II	57 (7.9)	13 (7.2)
III	148 (20.5)	36 (19.9)
IV	468 (64.8)	120 (66.3)
Unknown	10 (1.4)	2 (1.1)
Site of metastasis or recurrence*		
None	59 (8.2)	18 (9.9)
Liver	160 (22.2)	38 (21.0)
Lung	231 (32.0)	54 (29.8)
Bone	122 (16.9)	38 (21.0)
Brain	71 (9.8)	27 (14.9)
Lymph nodes	284 (39.3)	77 (42.5)
Others	182 (25.2)	43 (23.8)
History of surgery		
Yes	396 (54.8)	84 (46.4)
No	325 (45.0)	97 (53.6)
Unknown	1 (0.1)	0 (0)
Type of treatment*		
Chemotherapy	469 (65.0)	113 (62.4)
Endocrine therapy	50 (6.9)	15 (8.3)
Molecular-targeted therapy	129 (17.9)	30 (16.6)
Immunotherapy	85 (11.8)	23 (12.7)
Palliative therapy	27 (3.7)	8 (4.4)
Others	6 (0.8)	2 (1.1)

Note. Median (IQR) was reported for age, whereas frequency (%) was reported for other characteristics.
ECOG indicates Eastern Cooperative Oncology Group; IQR, interquartile range.
*Multiple choices were allowed.

mapping were physical functioning for mobility, physical functioning for self-care, role functioning for usual activities, pain for pain/discomfort, emotional functioning for anxiety/depression, and global health status for not full health. In not full health analysis, pain, role functioning, and fatigue had an importance of more than three-quarters relative to the global health status. The results of other predictor variable importance measures (the frequency and cover) were reported in Appendix Figs. 4 and 5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.07.020. The most importance predictor variable for each outcome was the same among the 3 measures of predictor variable importance except for not full health.

The R code to predict the EQ-5D-5L index using the developed GBT-based mapping algorithms is freely available at https://github.com/YasuhiroHagiwara/GBTmapping.

Predictive Performances of GBT and Regression Approaches

Figure 3 shows predictive performances in the training data set (numerical data are in Appendix Table 4 in the Supplemental Material found at https://doi.org/10.1016/j.jval.2022.07.020). The RMSE and MAE in the whole data set were smaller in the GBT approaches for both the Japanese and US value sets than in the regression approaches. All approaches resulted in a larger RMSE and MAE as health condition deteriorated (ie, the global health status score was smaller). From the results of the ME, the 2 regression approaches and the GBT approach for response systematically overpredicted the EQ-5D-5L index by more than 0.01 on average in at least one of the poor health subgroups (the global health status score < 50). The GBT approach for the index did not provide such overprediction. All approaches except the 2-part beta regression underpredicted the EQ-5D-5L index by more than 0.01 on average in the best health subgroup (the global health status score ≥ 90).

Figure 4 shows predictive performances in the test data set (numerical data are in Appendix Table 5 in the Supplemental Material found at https://doi.org/10.1016/j.jval.2022.07.020). The RMSE and MAE in the whole data set were smaller in the regression approaches for both the Japanese and US value sets than in the GBT approaches. Similar to the training data set, the RMSE and MAE were larger as health condition deteriorated (ie, the global health status score was smaller). In terms of the ME, all approaches overpredicted the EQ-5D-5L index by more than 0.01 in the poorest health condition (the global health status score $<$ 30) but the degree of overprediction was the smallest in the GBT approach for index of the Japanese value set. Underprediction of the EQ-5D-5L index in the best health condition (the global health status score ≥ 90) was less than 0.01 in the GBT for index and response and 2-part beta regression for the Japanese value set.

Discussion

In this study, we developed direct and response mapping algorithms for the EORTC QLQ-C30 onto the EQ-5D-5L index using GBT approaches. The GBT-based mapping algorithms provided larger RMSE and MAE in the test data set than regression-based mapping algorithms. Overprediction in good health and underprediction in poor health tended to be smaller using the GBT approach for the index than the regression-based approaches. We provided the R code on GitHub (https://github.com/YasuhiroHagiwara/GBTmapping) to generate the mapped EQ-5D-5L index from the EORTC QLQ-C30 based on GBT and regression approaches in health technology assessment and health economic evaluation.[21,22]

The prediction error of the GBT-based direct mapping algorithms was the smallest among the evaluated mapping algorithms in the training data set, and this algorithm decreased overprediction in poor health and underprediction in good health compared with the regression-based mapping algorithms. This result reflects the flexibility of GBTs, but this flexibility induced overfitting; in the test data set, the prediction error measured by the RMSE and MAE was larger using the GBT-based direct mapping algorithms than the regression-based mapping algorithms. Overall, the GBT approach for direct mapping was helpful in

Table 2. Distributions of the EQ-5D-5L index and scores of EORTC QLQ-C30 in the training data set and the test data set.

	Minimum	5%	25%	Median	75%	95%	Maximum
Training data set							
EQ-5D-5L index							
Japan	−0.025	0.428	0.670	0.823	0.895	1	1
United States	−0.573	0.226	0.669	0.844	0.940	1	1
EORTC QLQ-C30							
Physical functioning	0	33.3	66.7	86.7	93.3	100	100
Role functioning	0	0	50	66.7	100	100	100
Emotional functioning	0	41.7	75	83.3	100	100	100
Cognitive functioning	0	33.3	66.7	83.3	100	100	100
Social functioning	0	33.3	66.7	83.3	100	100	100
Global health status	0	16.7	50	66.7	83.3	100	100
Fatigue	0	0	22.2	33.3	55.6	88.9	100
Nausea and vomiting	0	0	0	0	16.7	33.3	100
Pain	0	0	0	16.7	33.3	66.7	100
Dyspnea	0	0	0	33.3	33.3	100	100
Insomnia	0	0	0	33.3	33.3	66.7	100
Appetite loss	0	0	0	0	33.3	100	100
Constipation	0	0	0	33.3	33.3	66.7	100
Diarrhea	0	0	0	0	33.3	66.7	100
Financial difficulties	0	0	0	33.3	33.3	100	100
Test data set							
EQ-5D-5L index							
Japan	0.034	0.491	0.732	0.817	0.895	1	1
United States	−0.477	0.341	0.719	0.836	0.940	1	1
EORTC QLQ-C30							
Physical functioning	0	40	66.7	83.3	93.3	100	100
Role functioning	0	16.7	66.7	83.3	100	100	100
Emotional functioning	8.3	50	75	83.3	100	100	100
Cognitive functioning	16.7	50	66.7	83.3	83.3	100	100
Social functioning	0	33.3	66.7	83.3	100	100	100
Global health status	0	16.7	41.7	66.7	83.3	100	100
Fatigue	0	0	22.2	33.3	44.4	77.8	100
Nausea and vomiting	0	0	0	0	0	50	100
Pain	0	0	0	16.7	33.3	66.7	100
Dyspnea	0	0	0	33.3	33.3	66.7	100
Insomnia	0	0	0	33.3	33.3	66.7	100
Appetite loss	0	0	0	33.3	33.3	100	100
Constipation	0	0	0	33.3	33.3	100	100
Diarrhea	0	0	0	0	33.3	66.7	100
Financial difficulties	0	0	0	33.3	33.3	100	100

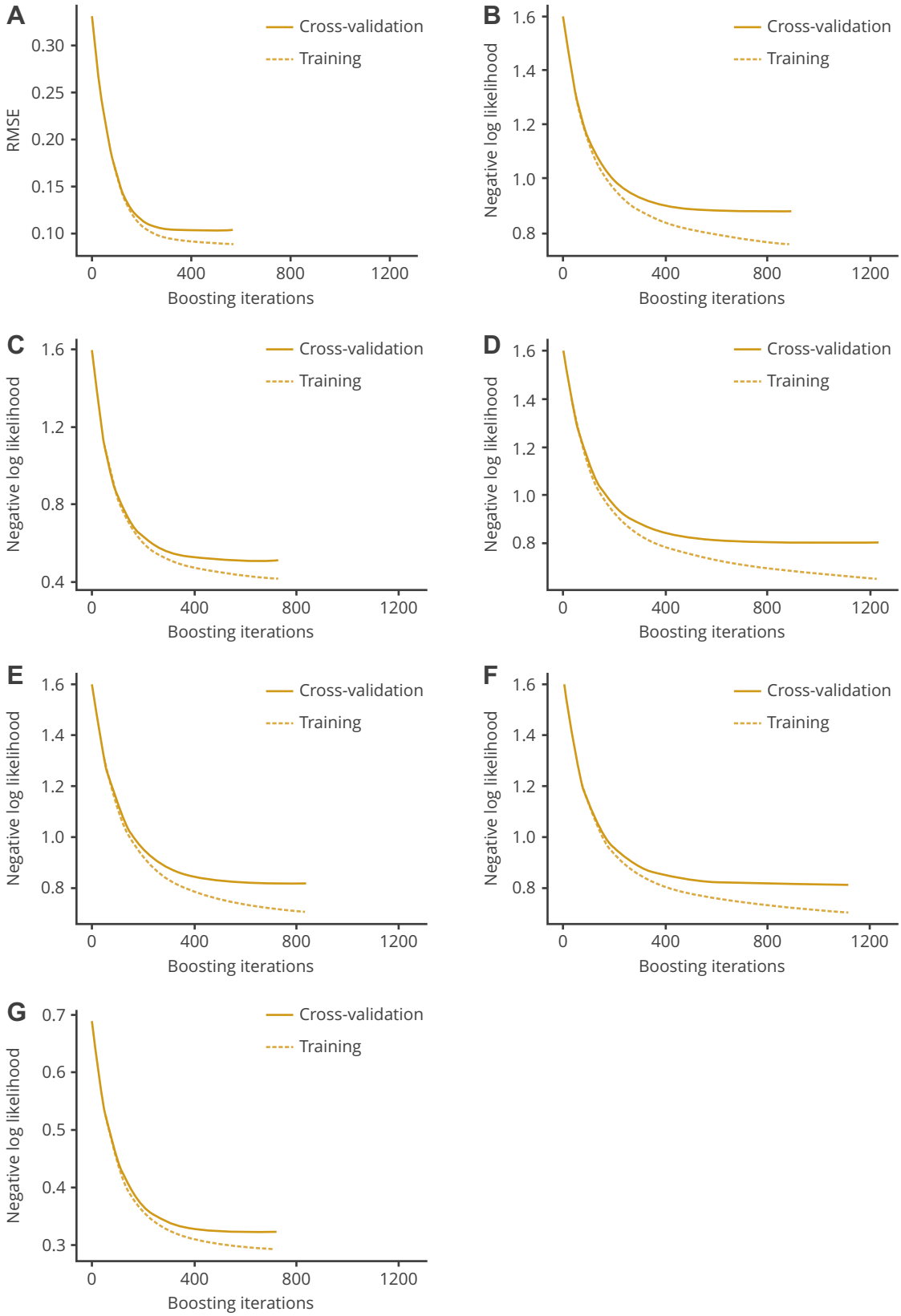
EORTC QLQ-C30 indicates European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30; EQ-5D-5L, 5-level version of EQ-5D.

reducing overprediction and underprediction, but it was not useful for generating the EQ-5D-5L index close to the true index compared with regression-based approaches.

Although the GBT approach was not useful for developing a more accurate response mapping algorithm than the regression-based approach in the test data set in terms of RMSE, MAE, and ME, the results suggest several important findings on the association between responses to the EQ-5D-5L and the EORTC QLQ-C30. First, a tree with a depth of 1 was selected in the cross-validation for all 5 items and not full health of the EQ-5D-5L. Given that the tree with a depth of 1 has 2 leaves split by 1 predictor variable, it indicates that interaction between scores in the EORTC QLQ-C30 did not improve the prediction of response to each item of the EQ-5D-5L and not full health. Second, only the subscale score in the EORTC QLQ-C30 that corresponds to each item of the EQ-5D-5L dominated other subscale scores in the predictor variable importance (eg, physical functioning for mobility). Third, because GBTs with a depth of 1, which can incorporate nonlinearity, did not have better predictive performance than ordinal logistic regression, it is also suggested that a nonlinear relation (in the logit scale) was not important in predicting response to each item of the EQ-5D-5L. These findings together suggest that the linear terms of a subscale score in EORTC QLQ-C30 that has substantial overlap to each item of EQ-5D-5L and a few other complementary subscale scores would be sufficient to accurately predict each dimension of the EQ-5D-5L in mapping from the EORTC QLQ-C30 with regression models.

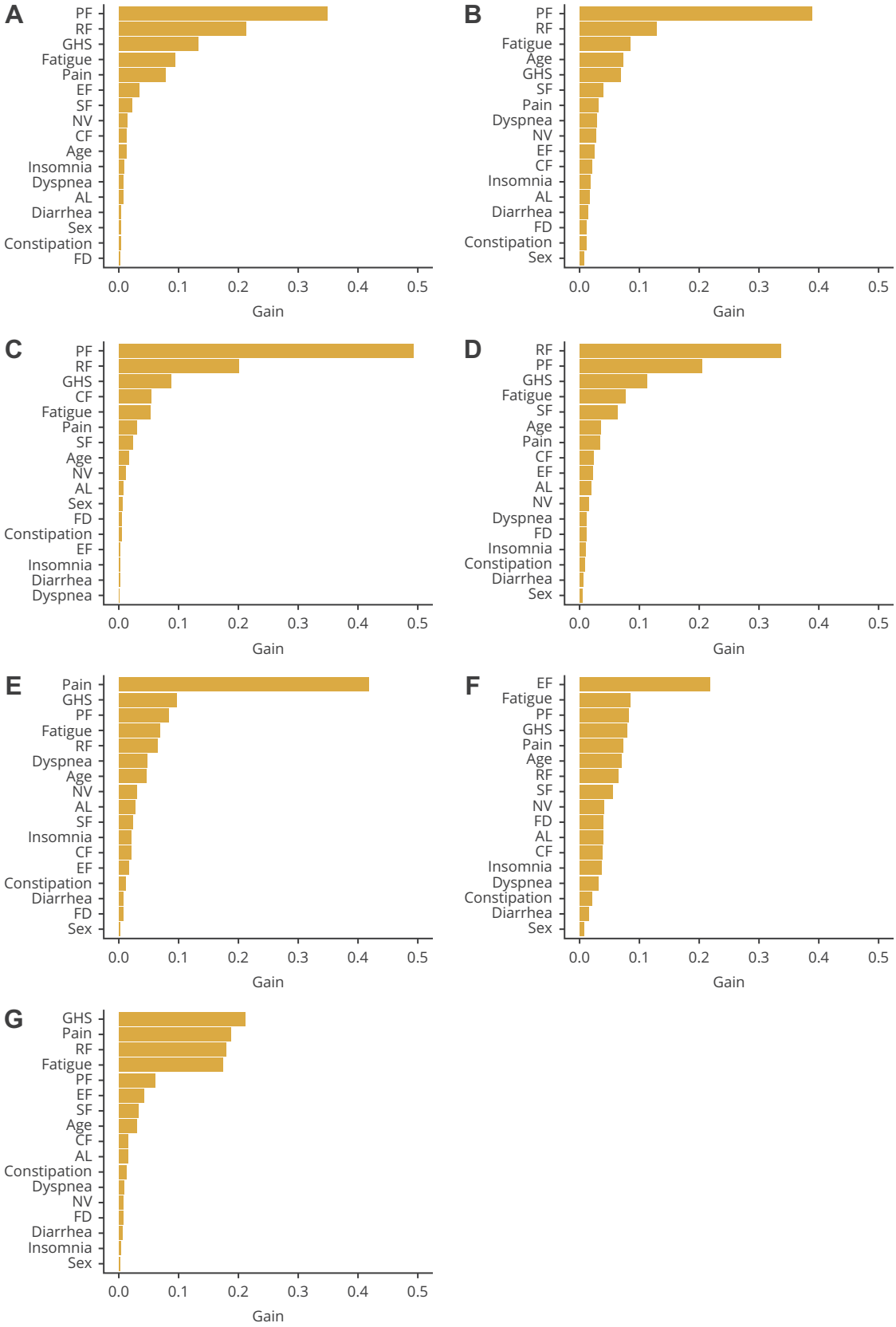
Nine previous studies developed mapping algorithms from the EORCT QLQ-C30 onto the EQ-5D-5L index.[23-31] Fair comparisons of the present study with all previous studies are not feasible due to different patient populations and value sets. Only 1 previous study evaluated predictive performances of mapping algorithms using the Japanese and US value sets.[31] In this previous study, the cross-validated RMSE by the best direct and response mapping algorithms for the Japanese value set was 0.099 by 2-part truncated ordinary least-squares regression and 0.098 by ordinal logistic regression, respectively, whereas that by the response mapping algorithm for the US value set was 0.150 by ordinal logistic regression. Compared with these results, no substantial improvement in overall predictive performances was observed with the GBT approaches in the present study.

Figure 1. Loss functions of GBTs over boosting iterations in the training and validation set in cross-validation. Boosting iterations were terminated after 100 consecutive failures to improve loss functions in the validation set. (A) RMSE in the EQ-5D-5L index analysis. (B) Negative multinomial log likelihood in the mobility analysis. (C) Negative multinomial log likelihood in the self-care analysis. (D) Negative multinomial log likelihood in the usual activities analysis. (E) Negative multinomial log likelihood in the pain/discomfort analysis. (F) Negative multinomial log likelihood in the anxiety/depression analysis. (G) Negative binary log likelihood in the not full health analysis.



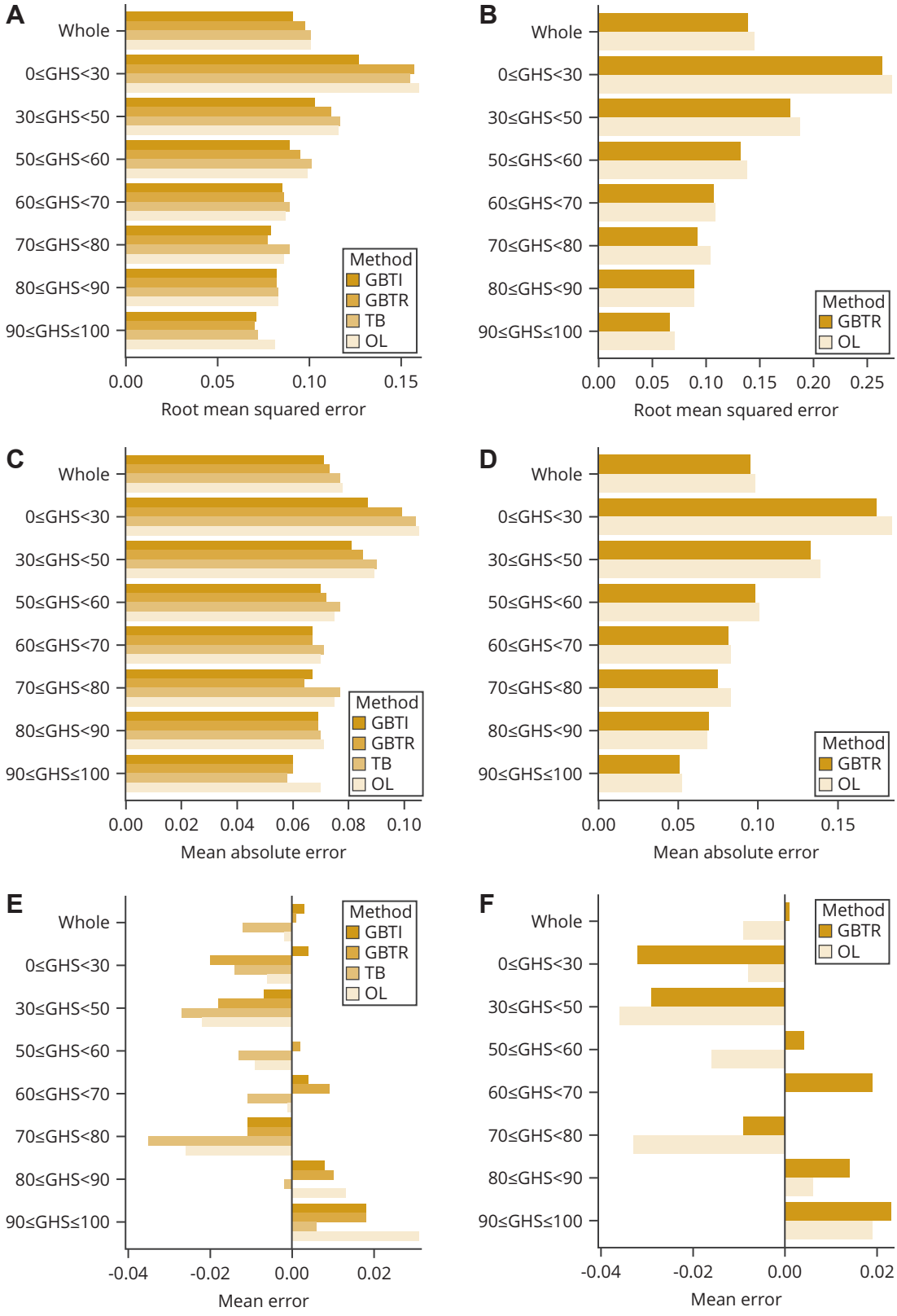
EQ-5D-5L indicates 5-level version of EQ-5D; GBT, gradient boosted tree; RMSE, root mean squared error.

Figure 2. Predictor variable importance measured by the gain in GBT analysis. The gain represents relative contribution of each variable to reducing a loss function. (A) EQ-5D-5L index analysis. (B) Mobility analysis. (C) Self-care analysis. (D) Usual activities analysis. (E) Pain/discomfort analysis. (F) Anxiety/depression analysis. (G) Not full health analysis.



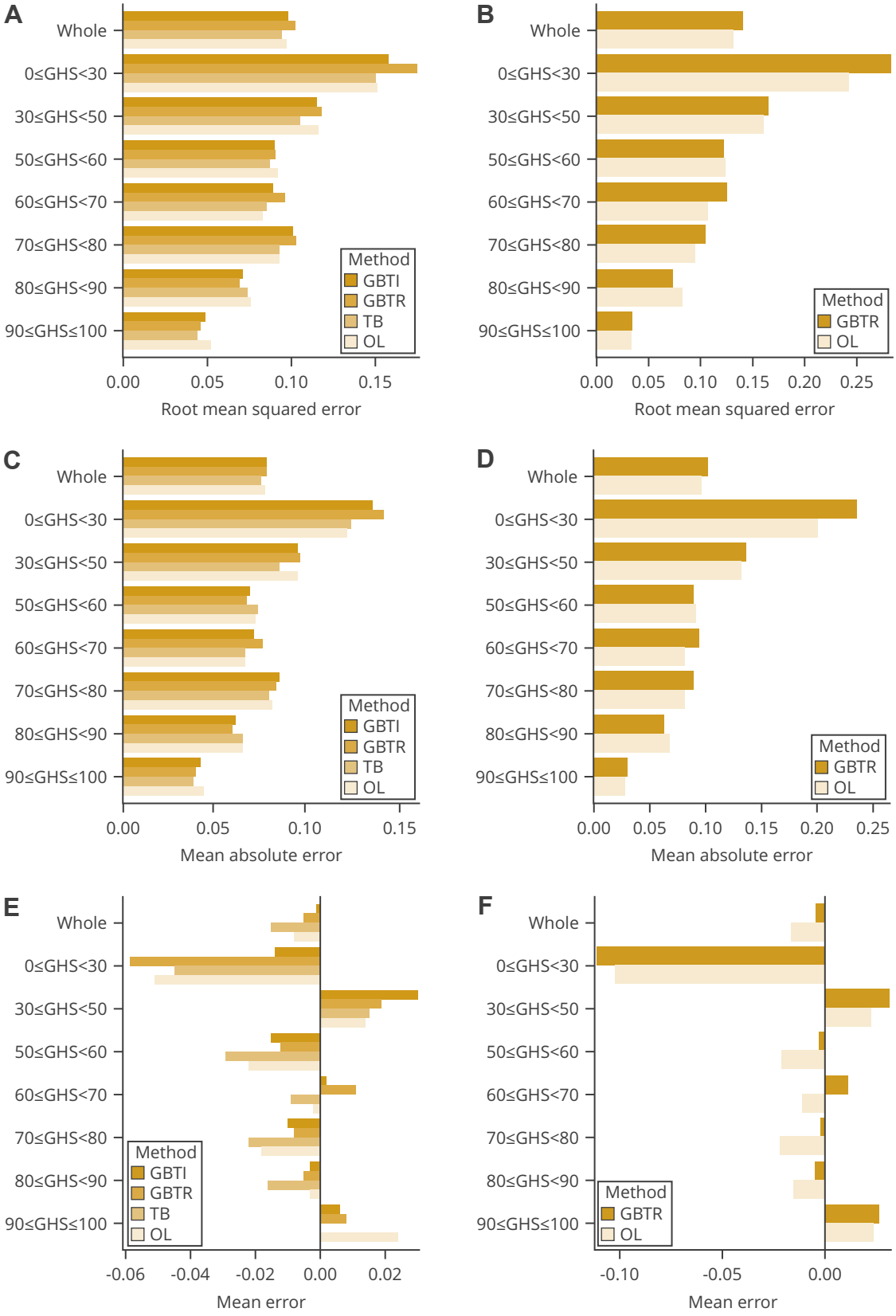
AL indicates appetite loss; CF, cognitive functioning; EF, emotional functioning; EQ-5D-5L, 5-level version of EQ-5D; FD, financial difficulties; GBT, gradient boosted tree; GHS, global health status; NV, nausea and vomiting; PF, physical functioning; RF, role functioning; SF, social functioning.

Figure 3. Predictive performance in the training dataset. (A) Root mean squared error for the Japanese value set. (B) Root mean squared error for the US value set. (C) Mean absolute error for the Japanese value set. (D) Mean absolute error for the US value set. (E) Mean error for the Japanese value set. (F) Mean error for the US value set.



GBTI indicates gradient boosted tree for index; GBTR, gradient boosted trees for response; GHS, global health status; OL, ordinal logistic regression; TB, 2-part beta regression.

Figure 4. Predictive performance in the test dataset. (A) Root mean squared error for the Japanese value set. (B) Root mean squared error for the US value set. (C) Mean absolute error for the Japanese value set. (D) Mean absolute error for the US value set. (E) Mean error for the Japanese value set. (F) Mean error for the US value set.



GBTI indicates gradient boosted tree for index; GBTR, gradient boosted trees for response; GHS, global health status; OL, ordinal logistic regression; TB, 2-part beta regression.

One challenge in mapping algorithms based on machine learning methods is how to account for the uncertainty of mapping algorithms and individual variability. In regression-based mapping algorithms, the uncertainty of mapping algorithms can usually be represented with a variance-covariance matrix of estimated coefficients, and individual variability can be dealt with by a parametric distribution assumed in a regression model.[3,32] Nevertheless, GBT does not have estimated coefficients, and hence, their variance-covariance matrix and the GBT approach for the index do not assume any parametric distribution. These properties are applicable to other modern machine learning methods, such as deep neural networks.[8] Future research should be conducted to incorporate methods that quantify uncertainty in prediction by GBTs into mapping from a nonpreference-based measure onto health utility.[33]

A major limitation of this research is the limited sample size. This limitation is relevant to both training GBTs and testing their predictive performances. If more data are used to train GBTs, they may provide more accurate predictions than regression approaches. In general, flexible machine learning methods require more data than traditional regression methods to perform well. The limited sample size in the test data set resulted in imprecise predictive performance estimates. Given that the differences in predictive performances between the 2 approaches were small, a definitive conclusion on which approach is best suited to mapping from EORTC QLQ-C30 onto EQ-5D-5L index could not be derived. To reliably investigate usefulness of machine learning methods in mapping from a nonpreference-based measure onto health utility, further research with a larger data set is needed, although we used the largest data set for mapping from EORTC QLQ-C30 onto EQ-5D-5L index for patients with solid tumors.[23–28,30,31]

Additional limitations should be considered when interpreting the results of this research. First, our results were based on 1 pair of a nonpreference-based measure (EORTC QLQ-C30) and a health utility measure (EQ-5D-5L). Other pairs of nonpreference-based measures and health utility measures may have complex association patterns, including nonlinear relations and interactions. In this case, a flexible machine learning method may substantially outperform existing regression-based mapping algorithms. Second, the regression-based mapping algorithms evaluated here were developed using all data from the QOL-MAC study (including the test data set) in the previous research.[12] Therefore, the comparison between the GBT and regression approaches in the test data set was somewhat in favor of the regression approaches because the regression approaches used data included in the test data set in training. An advantage of the regression approach is that it does not require a test data set given that the predictive performance can be evaluated in cross-validation, whereas the GBT approaches use cross-validation for tuning hyperparameters and hence require a test data set. Nevertheless, this effect would be small because the degree of overfitting of the regression-based mapping algorithm was negligible.[12] Therefore, it is unlikely that the developed GBT-based mapping algorithms substantially outperformed the regression-based mapping algorithms in a fair comparison.

Conclusions

We developed direct and response mapping algorithms for the EORTC QLQ-C30 onto the EQ-5D-5L index based on GBTs. Compared with regression-based mapping algorithms developed from the same data, GBT-based mapping algorithms did not improve the predictive performance measured by the RMSE and MAE in the test data set but had the potential to reduce overprediction in poor health and underprediction in good health. Further research is needed to establish the role of modern flexible machine learning methods in mapping a nonpreference-based measure onto health utility.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.07.020.

Article and Author Information

Accepted for Publication: July 31, 2022

Published Online: xxxx

doi: https://doi.org/10.1016/j.jval.2022.07.020

Author Affiliations: Department of Biostatistics, Division of Health Sciences and Nursing, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan (Hagiwara); Center for Outcomes Research and Economic Evaluation for Health, National Institute of Public Health, Wako, Japan (Shiroiwa, Konomura, Fukuda); Department of Breast and Thyroid Surgery, Kawasaki Medical School, Kurashiki, Japan (Taira); Clinical Research Promotion Center, The University of Tokyo Hospital, Tokyo, Japan (Kawahara); Center for Health Economics and QOL Research, Niigata University of Health and Welfare, Niigata, Japan (Noto); Department of Biomedical Sciences, College of Life Sciences, Ritsumeikan University, Kusatsu, Japan (Shimozuma).

Correspondence: Yasuhiro Hagiwara, PhD, MPH, Department of Biostatistics, Division of Health Sciences and Nursing, Graduate School of Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. Email: hagiwara@epistat.m.u-tokyo.ac.jp

Author Contributions: *Concept and design:* Hagiwara, Shiroiwa, Taira, Kawahara, Konomura, Noto, Fukuda, Shimozuma
Acquisition of data: Taira
Analysis and interpretation of data: Hagiwara, Kawahara
Drafting of the manuscript: Hagiwara
Critical revision of the paper for important intellectual content: Hagiwara, Shiroiwa, Taira, Kawahara, Konomura, Noto, Fukuda, Shimozuma
Statistical analysis: Hagiwara
Provision of study materials or patients: Taira, Noto
Administrative, technical, or logistic support: Konomura
Supervision: Shiroiwa, Fukuda, Shimozuma

Conflict of Interest Disclosures: Dr Hagiwara reported receiving personal fees from SAS Institute Japan, outside the submitted work. Dr Taira reported receiving personal fees from Pfizer Inc, Eisai Co, Ltd, Chugai Pharma, and Kyowa Kirin Co, Ltd, outside the submitted work. No other disclosures were reported.

Funding/Support: The QOL-MAC study was sponsored by the Public Health Research Foundation (PHRF). The research fund was provided to the PHRF by the Center for Outcomes Research and Economic Evaluation for Health, National Institute of Public Health, under the study contract. This work was partially supported under grant 19K24193 from the Japan Society for the Promotion of Science, KAKENHI.

Role of Funder/Sponsor: Researchers at the Center for Outcomes Research and Economic Evaluation for Health, National Institute of Public Health (Shiroiwa, Konomura, and Fukuda), participated in the study design, data collection, data analysis, data interpretation, and the writing of the report.

Acknowledgment: The authors thank the patients who participated in the QOL-MAC study. The QOL-MAC study was supported by the Comprehensive Support Project for Oncology Research of the PHRF.

REFERENCES

1. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *Eur J Health Econ.* 2010;11(2):215–225.

2. Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in nice health technology assessments. *Value Health*. 2013;16(1):202–210.
3. Wailoo AJ, Hernandez-Alava M, Manca A, et al. Mapping to estimate health-state utility from non–preference-based outcome measures: an ISPOR good practices for outcomes research task force report. *Value Health*. 2017;20(1):18–27.
4. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? *Health Qual Life Outcomes*. 2009;7:27.
5. Versteegh MM, Rowen D, Brazier JE, Stolk EA. Mapping onto EQ-5D for patients in poor health. *Health Qual Life Outcomes*. 2010;8:141.
6. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer-Verlag; 2009.
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–1358.
8. Gao L, Luo W, Tonmukayakul U, Moodie M, Chen G. Mapping MacNew Heart Disease Quality of Life Questionnaire onto country-specific EQ-5D-5L utility scores: a comparison of traditional regression models with a machine learning technique. *Eur J Health Econ*. 2021;22(2):341–350.
9. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232.
10. Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. In: Hastie T, Tibshirani R, Friedman J, eds. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer-Verlag; 2009: 337–388.
11. Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. *Value Health*. 2019;22(7):808–815.
12. Hagiwara Y, Shiroiwa T, Taira N, et al. Mapping EORTC QLQ-C30 and FACT-G onto EQ-5D-5L index for patients with cancer. *Health Qual Life Outcomes*. 2020;18(1):354.
13. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–1736.
14. Shiroiwa T, Ikeda S, Noto S, et al. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. *Value Health*. 2016;19(5):648–654.
15. Pickard AS, Law EH, Jiang R, et al. United States valuation of EQ-5D-5L health states using an international protocol. *Value Health*. 2019;22(8):931–941.
16. 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(5):365–376.
17. James G, Witten D, Hastie T, Tibshirani R. Tree-based methods. In: James G, Witten D, Hastie T, Tibshirani R, eds. *An Introduction to Statistical Learning With Applications in R*. New York, NY: Springer; 2013:303–335.
18. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367–378.
19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, CA; August 13-17, 2016. 785-794. https://dl.acm.org/doi/10.1145/2939672.2939785.
20. Chen T, He T, Benesty M, et al. xgboost: extreme gradient boosting. https://cran.r-project.org/web/packages/xgboost/index.html. Accessed November 19, 2021.
21. Incerti D, Thom H, Baio G, Jansen JP. R you still using Excel? The advantages of modern software tools for health technology assessment. *Value Health*. 2019;22(5):575–579.
22. Hart R, Burns D, Ramaekers B, et al. R and Shiny for cost-effectiveness analyses: why and when? A hypothetical case study. *Pharmacoeconomics*. 2020;38(7):765–776.
23. Khan I, Morris S, Pashayan N, Matata B, Bashir Z, Maguirre J. Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients. *Health Qual Life Outcomes*. 2016;14:60.
24. Lamu AN, Olsen JA. Testing alternative regression models to predict utilities: mapping the QLQ-C30 onto the EQ-5D-5L and the SF-6D. *Qual Life Res*. 2018;27(11):2823–2839.
25. Ameri H, Yousefi M, Yaseri M, Nahvijou A, Arab M, Akbari Sari A. Mapping the cancer-specific QLQ-C30 onto the generic EQ-5D-5L and SF-6D in colorectal cancer patients. *Expert Rev Pharmacoecon Outcomes Res*. 2019;19(1):89–96.
26. Ameri H, Yousefi M, Yaseri M, Nahvijou A, Arab M, Akbari Sari A. Mapping EORTC-QLQ-C30 and QLQ-CR29 onto EQ-5D-5L in colorectal cancer patients. *J Gastrointest Cancer*. 2020;51(1):196–203.
27. Noel CW, Stephens RF, Su JS, et al. Mapping the EORTC QLQ-C30 and QLQ-H&N35, onto EQ-5D-5L and HUI-3 indices in patients with head and neck cancer. *Head Neck*. 2020;42(9):2277–2286.
28. Liu T, Li S, Wang M, Sun Q, Chen G. Mapping the Chinese version of the EORTC QLQ-BR53 onto the EQ-5D-5L and SF-6D utility scores. *Patient*. 2020;13(5):537–555.
29. Xu RH, Wong ELY, Jin J, Dou Y, Dong D. Mapping of the EORTC QLQ-C30 to EQ-5D-5L index in patients with lymphomas. *Eur J Health Econ*. 2020;21(9):1363–1373.
30. Yousefi M, Nahvijou A, Sari AA, Ameri H. Mapping QLQ-C30 Onto EQ-5D-5L and SF-6D-V2 in patients with colorectal and breast cancer from a developing country. *Value Health Reg Issues*. 2021;24:57–66.
31. Meunier A, Soare A, Chevrou-Severac H, Myren K-J, Murata T, Longworth L. Indirect and direct mapping of the cancer-specific EORTC QLQ-C30 onto EQ-5D-5L utility scores. *Appl Health Econ Health Policy*. 2022;20(1):119–131.
32. Petrou S, Rivero-Arias O, Dakin H, et al. The MAPS reporting statement for studies mapping onto generic preference-based outcome measures: explanation and elaboration. *Pharmacoeconomics*. 2015;33(10):993–1011.
33. Malinin A, Prokhorenkova L, Ustimenko A. Uncertainty in gradient boosting via ensembles. arXiv. https://doi.org/10.48550/arXiv.2006.10562.