ScienceDirect

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

Patient-Reported Outcomes

Predicting Patient-Level 3-Level Version of EQ-5D Index Scores From a Large International Database Using Machine Learning and Regression Methods

Zsombor Zrubka, MD, PhD, István Csabai, DSc, Zoltán Hermann, PhD, Dominik Golicki, MD, PhD, Valentina Prevolnik-Rupel, PhD, Marko Ogorevc, PhD, László Gulácsi, MD, DSc, Márta Péntek, MD, PhD

ABSTRACT

Objectives: This study aimed to evaluate the performance of machine learning and regression methods in the prediction of 3-level version of EQ-5D (EQ-5D-3L) index scores from a large diverse data set.

Methods: A total of 30 studies from 3 countries were combined. Predictions were performed via eXtreme Gradient Boosting classification (XGBC), eXtreme Gradient Boosting regression (XGBR) and ordinary least squares (OLS) regression using 10-fold cross-validation and 80%/20% partition for training and testing. We evaluated 6 prediction scenarios using 3 samples (general population, patients, total) and 2 predictor sets: demographic and disease-related variables with/without patient-reported outcomes. Model performance was evaluated by mean absolute error and percent of predictions within clinically irrelevant error range and within correct health severity group (EQ-5D-3L index <0.45, 0.45-0.926, >0.926).

Results: The data set involved 26 318 individuals (clinical settings n = 6214, general population n = 20 104) and 26 predictor variables plus diagnoses. Using all predictors and the total sample, mean absolute error values were 0.153, 0.126, and 0.131, percent of predictions within clinically irrelevant error range were 47.6%, 39.5%, and 37.4%, and within the correct health severity group were 56.3%, 64.9%, and 63.3% by XGBC, XGBR, and OLS, respectively. The performance of models depended on the applied evaluation criteria, the target population, the included predictors, and the EQ-5D-3L index score range.

Conclusions: Regression models (XGBR and OLS) outperformed XGBC, yet prediction errors were outside the clinically irrelevant error range for most respondents. Our results highlight the importance of systematic patient-reported outcome (EQ-5D) data collection. Dialogs between artificial intelligence and outcomes research experts are encouraged to enhance the value of accumulating data in health systems.

Keywords: EQ-5D-3L, machine learning, prediction.

VALUE HEALTH. 2022; 25(9):1590–1601

Introduction

Patient-reported outcomes (PROs) reflect disease burden or treatment effectiveness from the patients' perspective. The value of PROs in improving health system performance and individual health outcomes has been demonstrated in multiple settings.[1] Preference-based health measures such as the EQ-5D are widely used in health economic evaluations.[2-5]

Although the monitoring of PROs has become a priority in many health systems, their organized collection at national level is still in its infancy.[1] With the gradual implementation of electronic health records and harmonized statistical data collections (eg, European Health Survey), a large amount of administrative health data is being collected.[6-8] Smart devices, big data, and advanced analytic techniques are contributing to the personalization of healthcare.[9-13] Nevertheless, because of varying rules of data sharing, standards of interoperability, available infrastructure or level of stakeholder collaboration, and data sets, which are usually collected at different time points for different purposes with different methods in the health data ecosystem, are difficult to connect.[14] For example, the Minimum European Health Module is a PRO measure collected regularly in Eurostat population surveys while hardly used in clinical trials. The EQ-5D questionnaire has been increasingly used in clinical trials, health surveys, and registries,[15,16] but infrequently in general clinical practice or administrative health surveys.[17,18] The accumulating big data are typically unstructured, heterogenous, and incomplete, which may hamper the analysis using standard regression methods, whereas novel machine learning (ML) approaches may offer advantages in such data sets.

For calculating quality-adjusted life-years in health economic analyses, EQ-5D values are often missing and have to be estimated

from other health measures.[19-21] Therefore, the question arises whether EQ-5D index scores can be predicted from a large diverse data set combined from multiple sources and whether novel analytical methods offer advantages over conventional regression techniques.

Over the past 15 years, we collected EQ-5D-3L data in 30 studies from 26 318 individuals in a variety of settings and designs.[22-52] As a model of the heterogenous sociodemographic and disease-related variables that can be yielded from real-world electronic health records, a combined anonymous data set was created by applying uniform data-management rules for standard sociodemographic and healthcare-related variables.

This study aimed to evaluate the performance of ML and ordinary least squares (OLS) regression in the prediction of individual-level EQ-5D-3L index scores from variables routinely collected in observational studies in various patient populations and the general population.

Methods

EQ-5D-3L

The EQ-5D-3L questionnaire consists of 2 parts. The descriptive system assesses self-reported health in 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. In each dimension, respondents can describe their current health with one of the following 3 categories: no problems, some problems, and severe problems. The descriptive system defines 243 (3^5) distinct health states.[5] The EQ-5D-3L index scores (utilities) attached to each health state are measured in valuation studies and reflect societal preferences. In this study, we applied the UK EQ-5D-3L index value set (range −0.594 to 1.000).[49] The EQ-5D-3L index score of 1 represents perfect health, 0 represents death, and negative values represent "worse than death" health states. The second part of the instrument is a 20-cm vertical EuroQol visual analog scale (EQ-VAS) for the measurement of current health ranging from 0 (worst imaginable health) to 100 (best imaginable health).

Study Population

Data were collected in Hungary, Poland, and Slovenia. These countries have EQ-5D-3L value sets[39,44,50] and population norms.[40,45,50,53] Between 2000 and 2015, nearly three-quarters of EQ-5D-related studies in Central and Eastern Europe originated from these 3 countries.[17] From Hungary, we involved 2421 outpatients with 18 chronic conditions including psoriatic arthritis (n = 177),[22] plaque psoriasis (PP, n = 192),[23] peripheric arterial occlusive disease (n = 103),[24] age-related macular degeneration (n = 122),[41] rheumatoid arthritis (n = 249),[25] systemic sclerosis (SSC, n = 80),[26] dementia (n = 86),[27] diabetes mellitus (n = 264),[28] endometriosis (n = 79),[29] osteoporosis (n = 207),[30] adult attention deficit-hyperactivity disorder (n = 75),[31] urinary bladder cancer (n = 148),[32] benign prostatic hyperplasia (BPH, n = 237),[33] epilepsy (n = 96),[34] overactive bladder (n = 61),[35] Parkinson's disease (n = 99),[36] chronic schizophrenia (n = 78),[37] and multiple sclerosis (n = 68).[38] Furthermore, we included 14 442 individuals from general population studies including a large representative health survey (HHU, n = 2019),[52] the Hungarian EQ-5D-3L/5-level version of EQ-5D (EQ-5D-5L) valuation study (VHU, n = 1000)[39]), a survey about health expectations among visitors of the largest online news portal (EXP, n = 9142),[42] and a representative survey aiming to measure the monetary value of a quality-adjusted life-year in Europe (n = 2281).[40,43] From Poland, we included 504 patients from cohort studies involving measurements before, during, or after hospitalization because of stroke (cerebrovascular accident, n = 397)[46] and osteoporotic hip fracture (n = 107)[47] and 4704 patients from the general population including students (STU, n = 443),[48] respondents from the Polish EQ-5D-3L valuation study (VPL, n = 320),[44] and respondents from the Polish EQ-5D-3L population norms study (n = 3941.[45] From Slovenia, 3290 patients were included from a cohort study comparing the outcomes of various health programs across hospitals including conditions such as gonarthrosis, coxarthrosis, intervertebral disc disease, urinary incontinence, carpal tunnel syndrome, inguinal hernia, varicose veins, osteosynthesis removal, and shoulder lesions (PSI, n = 3290),[51] and 958 respondents from the general population, including the Slovenian EQ-5D-3L VAS valuation study (n = 734)[50] and the Slovenian EQ-5D-3L time trade-off valuation study (n = 224).[49] Three studies involved multiple measurements of EQ-5D-3L from the same patient (eg, before and after hospitalization).[46,47,51] The detailed description of the involved study populations is provided in the reference publications. We involved patients from the original databases without any other restrictions if EQ-5D-3L values were available; therefore, the number of eligible patients for our study may have differed slightly from the reference publications.

Database and Variables

We partitioned the database into a general population[39,40,42,44,45,48-50,52] and patient population[22-38,41,46,47,51] sample. The database structure is summarized by listing key study characteristics and nonmissing values for patient-level explanatory variables in Appendix Table 1 of Appendix 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024. The dependent variable was the EQ-5D-3L index score.[54] Patient-level predictor variables were organized into 2 groups.

Demographic and disease-related variables

Predictors in this group included age, gender, education, place of residence, family status, employment, relative income (net personal income as a percent of the study year's national average net income), setting (outpatient, hospitalized, and postoperative in the case of documented surgery within 3-6 months), number of general practitioner visits, any general practitioner visit, number of specialist visits, and any specialist visits or hospitalizations in the past year. In the case of missing data, we assumed that specialist visits happened for patients recruited in outpatient specialist centers,[23,24,41] and both specialist visits and hospitalizations happened at patients whose EQ-5D data were collected during or after hospitalization.[46,47,51] We recorded whether patients were informal care recipients, weight, height, and body mass index. Physician-reported outcomes were transferred to a standard scale, where 0 represents the worst and 1 the best possible health status measurable with the given instrument. The included physician-reported instruments (acronym; score of worst health status—score of best health status) were the clinician-reported VAS (0-100)[55] in the SSC and BPH studies,[26,33] the Mini-Mental State Examination (0-30)[56] in the dementia study,[27] the Clinical Global Impression (7-0) in the overactive bladder and schizophrenia studies,[35,37] the Expanded Disability Status Scale (10-0)[57] in the multiple sclerosis study,[38] and the Psoriasis Area and Severity Index (72-0)[58] in the PP study.[23]

Specific diseases were included as dummy variables indicating the main diagnosis of clinical populations[22-30,32-38,41,46,47,51,52,59] or self-reported conditions in 3 surveys in the general population (HHU,[52] VHU,[59] VPL[44]). An overall dummy variable indicated the presence of any disease. Appendix Table 2 of Appendix 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.

01.024 summarizes the number of patients in each study with conditions categorized under the International Classification of Diseases, Tenth Revision, codes and the International Classification of Diseases, Tenth Revision, main chapters.[60]

PRO

The second predictor group comprised data obtained from PROs such as EQ-VAS,[61] happiness measured on an 11-point (0-10) numeric scale,[62] and items of the Minimum European Health Module[63,64]: self-rated health and the limitations because of health problems (Global Activity Limitation Indicator). Scores from PRO instruments were transferred to a standard scale, with 0 representing the worst and 1 the best health status measurable with the given instrument. The applied PRO instruments (acronym; score of worst health status—score of best health status) were the Health Assessment Questionnaire Disability Index (3-0)[65] in the psoriatic arthritis, rheumatoid arthritis, and SSC studies[22,25,26]; the Dermatology Quality of Life Index (30-0)[66] in the PP study[23]; the Barthel Index (0-100)[67] in the cerebrovascular accident study[46]; the Functional Recovery Score (0%-100%)[68] in the hip fracture study[47]; the Bladder Cancer Index (0-100) in the BC study[69]; and the International Prostate Symptom Score (35-0)[70] in the BPH study.[33]

Data Analysis

Missing data

Missing data were handled via the indicator method. We imputed zeros for all missing values and generated a dummy indicator for each predictor denoting missing values. The indicator was set as missing in those general population studies, where self-reported conditions were not inquired. In contrast, the disease dummy was set as absent for those patients who were asked about the presence of a disease and responded negatively. Comorbidities were not recorded in patient populations; therefore, the disease dummy was set as missing except for the index conditions.

Prediction models

EQ-5D-3L index scores were predicted by OLS regression, eXtreme Gradient Boosting (XGBoost) classification (XGBC), and XGBoost regression (XGBR). A regular winner of ML competitions, XGBoost is a highly scalable and computationally efficient implementation of gradient boosted trees. Boosted decision trees are an ensemble of decision trees added sequentially. Each additional tree is trained to correct the errors of the ensemble of previous trees until no further improvements can be made on a validation data set. Gradient boosting grows the best trees by optimizing a loss function that comprises prediction error and a regularization term, which describes the complexity of the trees. Depending on the loss function, XGBoost can run in classification and regression mode, which predict EQ-5D scores in 243 categories or as a continuous value, respectively.[71,72]

Patients with multiple measurements were entered as unique records. No weights were applied. In the OLS model, age was split into 5-year categories, and a piecewise model was fit on EQ-VAS scores with different slopes for the 0 to 34 and 35 to 100 value ranges. We entered predictors without interactions but explored an OLS model with interaction between disease dummies, gender and age. For XGBoost, default settings were retained for most parameters after initial exploration and monitoring of train and test errors. The learning rate parameter was set to 0.1, number of trees were set to 20 to improve speed of classification, and the L1 regularization term of regression was set to 0.9 to avoid overfitting.

We performed predictions in 6 scenarios involving the general population sample ("Pop"), the patient sample ("Pts"), and the entire sample ("Total"), by using only demographic and disease predictors ("Base") and adding PRO predictors ("PRO"). Model training and evaluation were performed via 10-fold cross-validation, using a randomly selected 80% of the data set for training and 20% for testing. OLS coefficient estimates, XGBoost settings, and feature importance tables for the PRO scenarios are presented in Appendix 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024 (Appendix Tables 3-5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024).

Evaluation of Model Performance

Models were compared via the "*mean absolute error*" (MAE) of prediction as an intuitive and stabile measure when comparing scenarios with different sample sizes.[73] Furthermore, assuming that prediction errors smaller than 0.074, the minimum clinically important difference (MCD) of EQ-5D-3L, are barely noticeable,[74] yet greater errors in either direction are undesirable, we calculated the percent of predictions within the "clinically irrelevant error range" (eg, predictions within true ± MCD range). Third, we assessed prediction bias via observing mean prediction errors. Finally, we calculated the percent of predictions within the "correct health severity group." For this, according to the trimodal distribution of EQ-5D-3L, index values $>$ 0.926 (eg, 1-MCD) denoted "full health," values between 0.926 and 0.45 denoted "medium health," and values $<$ 0.45 denoted "low health."[75,76] (In "full health," true EQ-5D-3L index scores are equal to 1, whereas predictions can take values $>$ 0.926.)

To evaluate the reliability of predictions, from the 10 cross-validation sets, we calculated 95% confidence intervals for the evaluation metrics using the following formula:

$$95\%\ \mathrm{CI} = \overline{CV} \pm 1.96 * \frac{1}{\sqrt{k}}\mathrm{SD} \quad (1)$$

where $\overline{CV}$ is the mean and SD is the SD of the evaluation metric (eg, MAE) of the k cross-validation sets.[77]

Given that algorithmic bias is of particular concern in healthcare applications,[78] we evaluated whether individuals in different health statuses may be affected adversely because of prediction error. In addition to mean prediction error, we quantified the percent of predictions within clinically irrelevant error range and the percent of predictions within the correct health severity group across the full range of the true EQ-5D index. (Lower values denote greater risk of flawed predictions.) Second, assuming that decisions would be based on predicted and not the unknown true values, we evaluated bias and the probability of accurate predictions across the range of predicted EQ-5D-3L index values (eg, positive predictive value).[79]

Results

Sample Characteristics

The database contained 28 862 records of 26 318 individuals. Cross-sectional studies of the general population provided 20 104 records, whereas 8758 records were from cross-sectional and cohort studies involving 6214 patients (single measurement, n = 3753; 2 measurements, n = 2378; 3 measurements, n = 83). Most data originated from Hungary (16 862 records; 58.4%), followed by Slovenia (6507 records, 22.6%) and Poland (5493 records, 19.0%). Counting the diagnosis related dummies as one variable, the 28

Table 1. Summary of predictor variables and missing values.

Predictor group	Variable	Category	Sample		
			General population ("Pop")	Patients ("Pts")	Entire sample ("Total")
"Base"	Age	Mean	41.1	55.6	45.4
		SD	15.5	16.6	17.2
		Missing (%)	0.0	4.2	1.3
	Gender	Male (%)	54.6	50.4	53.3
		Female (%)	45.4	49.6	46.7
		Missing (%)	0.0	2.8	0.9
	Education	Primary (%)	10.8	23.3	14.2
		Secondary (%)	39.7	57.6	44.6
		Tertiary (%)	49.5	19.1	41.2
		Missing (%)	0.0	12.9	3.9
	Place of residence	Capital (%)	18.1	6.2	13.5
		City (%)	51.5	45.2	49.0
		Village (%)	30.4	48.6	37.5
		Missing (%)	56.8	36.6	50.7
	Family status	Single (%)	36.2	35.0	36.1
		Married (%)	63.8	65.0	63.9
		Missing (%)	26.6	77.0	41.9
	Employment	Paid employment (%)	68.6	42.9	61.6
		Student (%)	9.2	12.8	10.2
		Pensioner (%)	15.0	33.7	20.1
		Not working (%)	5.0	6.7	5.5
		Other employment (%)	2.2	3.9	2.6
		Missing (%)	0.0	13.4	4.1
	Relative income (0-11.0)	Mean	1.6	0.5	1.6
		SD	1.5	0.3	1.5
		Missing (%)	32.0	97.2	51.8
	Setting	General population (%)	100.0	-	69.7
		Outpatient (%)	-	30.1	9.1
		Hospitalized (%)	-	39.2	11.9
		Postoperative (%)	-	30.7	9.3
	Number of GP visits at 12 months	Mean	-	4.0	4.0
		SD	-	6.1	6.1
		Missing (%)	100.0	78.7	93.5
	Any GP visit past year	No (%)	-	49.1	49.1
		Yes (%)	-	50.9	50.9
		Missing (%)	100.0	78.7	93.5
	Specialist visits past year	Mean	-	5.8	5.8
		SD	-	7.5	7.5
		Missing (%)	100.0	80.9	94.2
	Any specialist visits past year	No (%)	-	17.1	17.1
		Yes (%)	-	82.9	82.9
		Missing (%)	100.0	80.9	94.2
	Hospitalizations past year	Mean	0.2	1.7	0.7
		SD	0.6	3.8	2.2
		Missing (%)	90.0	86.5	89.0

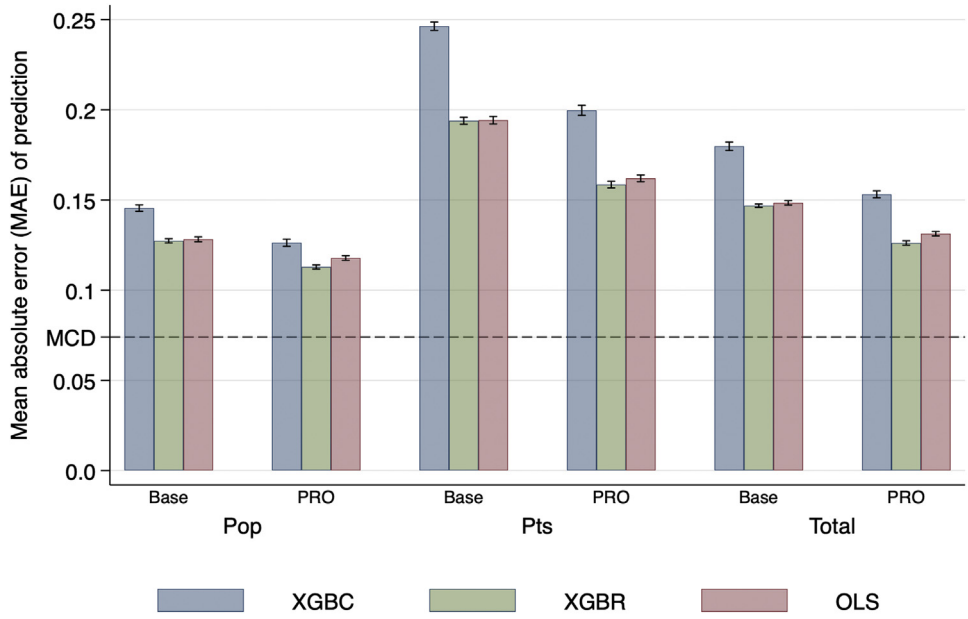
continued on next page

Table 1. Continued

Predictor group	Variable	Category	Sample		
			General population ("Pop")	Patients ("Pts")	Entire sample ("Total")
	Any hospitalization at 12 months	No (%)	90.4	41.9	72.5
		Yes (%)	9.6	58.1	27.5
		Missing (%)	90.0	86.5	89.0
	Informal care recipient	No (%)	92.2	70.9	82.4
		Yes (%)	7.8	29.1	17.6
		Missing (%)	90.0	80.3	87.1
	Weight, kg	Mean	76.1	75.2	75.6
		SD	16.1	16.8	16.4
		Missing (%)	88.9	72.8	84.0
	Height, cm	Mean	171.5	167.6	169.7
		SD	9.4	9.7	9.8
		Missing (%)	88.8	77.7	85.5
	BMI	Mean	25.8	26.8	26.3
		SD	4.8	5.2	5.0
		Missing (%)	88.9	77.8	85.5
	DRO score*	Mean	-	0.7	0.7
		SD	-	0.2	0.2
		Missing (%)	100.0	91.1	97.3
	Chronic morbidity	No (%)	68.5	0.0	68.5
		Yes (%)	31.5	0.0	31.5
		Missing (%)	90.0	100.0	93.0
	Any disease	No (%)	70.5	0.0	29.5
		Yes (%)	29.5	100.0	70.5
		Missing (%)	68.8	0.0	47.9
	Specific diagnoses*	Not included in the table			
"PRO"	Happiness	Mean	7.6	-	7.6
		SD	2.0	-	2.0
		Missing (%)	90.0	100.0	93.0
	Self-rated health	Very good (%)	20.7	0.0	20.7
		Good (%)	45.3	0.0	45.3
		Fair (%)	26.9	0.0	26.9
		Bad (%)	6.2	0.0	6.2
		Very Bad (%)	0.9	0.0	0.9
		Missing (%)	90.0	100.0	93.0
	GALI	Severely limited (%)	3.3	0.0	3.3
		Limited, but not severely (%)	16.8	0.0	16.8
		Not limited (%)	79.9	0.0	79.9
		Missing (%)	90.0	100.0	93.0
	PRO score*	Mean	-	0.7	0.7
		SD	-	0.3	0.3
		Missing (%)	100.0	80.7	94.1
	EQ-VAS (0-100)	Mean	77.0	65.4	73.3
		SD	18.9	22.3	20.7
		Missing (%)	11.5	6.3	9.9

Note. Mean, SD, and percent (%) values refer to nonmissing data. The percent of missing data was calculated for the entire sample.
BMI indicates body mass index; DRO, physician-reported outcome; EQ-VAS indicates EuroQol visual analog scale; GALI, Global Activity Limitation Indicator; GP, general practitioner; Pop, population; PRO, patient-reported outcome; Pts, patients.
*Details of PRO instruments and DRO instruments and disease dummies are omitted from the table.

Figure 1. MAE of predictions by scenario.



CI indicates confidence interval; MAE, mean absolute error; MCD, minimum clinically important difference; OLS, ordinary least squares; Pop, population; PRO, patient-reported outcome; Pts, patients; XGBC, eXtreme Gradient Boosting classification; XGBR, eXtreme Gradient Boosting regression.

predictor variables contained 64.1% missing values. There were 214 missing data patterns across observations with a range of missing variables from 7 to 23. There were no complete cases in the data set. The predictor variables and missing values are summarized in Table 1.

Mean (SD) EQ-5D-3L index scores were 0.847 (0.198), 0.665 (0.317), and 0.792 (0.254) in the general population, patients, and the total sample, respectively. In the general population sample, 3.7%, 49.0%, and 47.3%; in the patient sample, 14.2%, 62.3%, and 23.1%; and in the total sample, 6.9%, 53.2%, and 39.9% had EQ-5D-3L index scores in the low, medium, and full health categories, respectively.

The Distribution of True and Predicted EQ-5D-3L Index Scores

The distribution of XGBC predictions resembled the trimodal distribution of true EQ-5D scores, yet predictions of full health were more frequent. The distributions of XGBR and OLS were unimodal with left skew and a peak below full health (Appendix Fig. 1 of Appendix 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024). The range of full health across all scenarios was 23.1% to 47.2% for true EQ-5D-3L values, 43.4% to 93.9% for XGBC, 1.0% to 26.2% for XGBR, and 1.9% to 19.2% for OLS. XGBR predictions exceeded the value of 1 less frequently than those of OLS.

Accuracy of Predictions

In all scenarios, MAE was greatest for XGBC and lowest for XGBR followed closely by OLS. In the PRO scenario, MAE was 0.126, 0.113, and 0.118 in the population sample; 0.200, 0.159, and 0.162 in the patient sample; and 0.153, 0.126, and 0.131 in the total sample using XGBC, XGBR, and OLS, respectively. Adding PROs to demographic and disease-related predictors decreased MAE on average by 0.022 (Fig. 1).

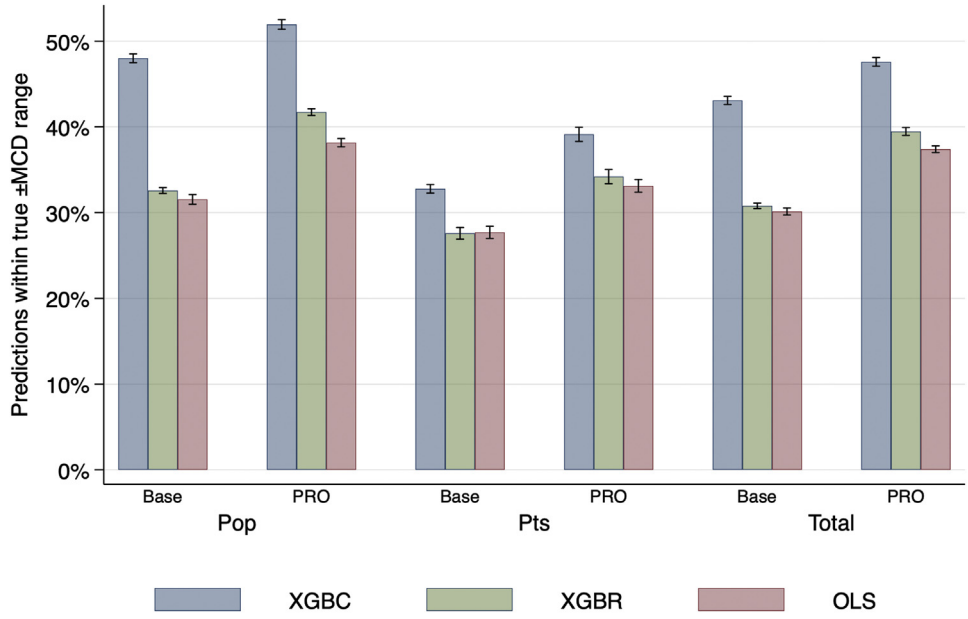
On the contrary, the percent of predictions within the clinically irrelevant error range (true ± MCD) was highest for XGBC with 51.9%, 39.1%, and 47.6%, followed by XGBR with 41.7%, 34.2%, and 39.5% and OLS with 38.2%, 33.1%, and 37.4% in the PRO scenario for the general population, patients, and the total sample, respectively. Adding PROs to the base predictors increased the percent of predictions within the clinically irrelevant error range on average by 6.6% (Fig. 2). Although mean prediction error for XGBR and OLS predictions was nearly zero in all scenarios, XGBC showed positive bias with mean error exceeding the MCD all scenarios (range 0.086-0.097) (Appendix Fig. 2 of Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024).

In terms of the percent of predictions within the correct health severity group, XGBR followed closely by OLS outperformed XGBC. In the PRO scenario for the general population, patient, and total samples, 57.2%, 58.1%, and 56.3% of predictions using XGBC; 63.2%, 68.5%, and 64.9% using XGBR; and 60.5%, 68.2%, and 63.1% using OLS fell within the correct health severity group, respectively (Fig. 3). The narrow 95% confidence interval ranges suggested that the predictions of all 3 methods were reliable through the cross-validation rounds. The performance of OLS models with or without interaction terms was rather similar (Appendix Fig. 3 of Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024).

Patterns of Prediction Error

XGBC often predicted full across the entire range of true EQ-5D-3L index values. The scatterplots of predicted over true values suggested that adding PROs to base predictors improved prediction accuracy mainly in the low health region and in

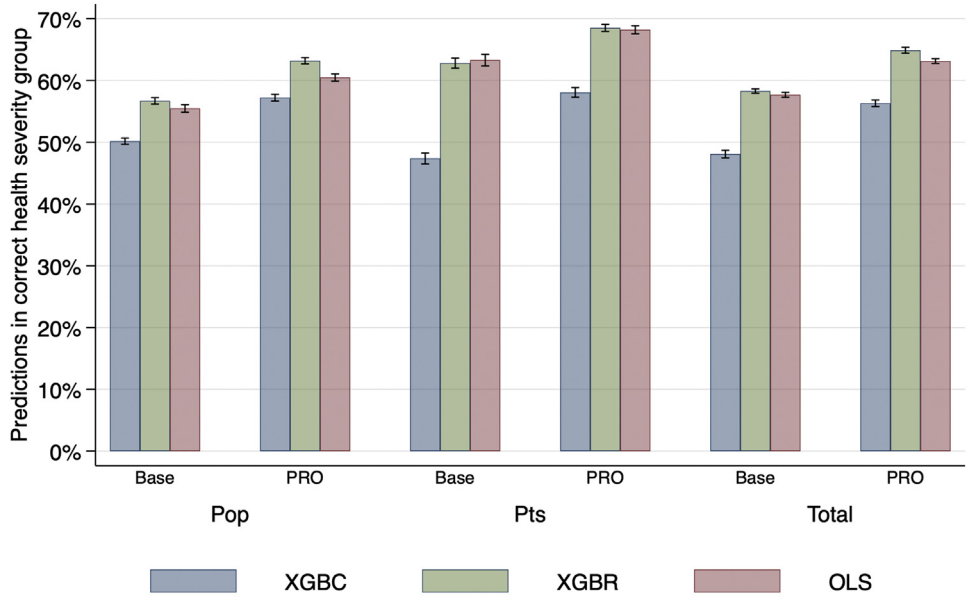
Figure 2. Percentage of predictions within clinically irrelevant error range (true ± MCD).



Error bars denote 95%CI of cross-validation; MCD: minimum clinically important difference

MCD indicates minimum clinically important difference; OLS, ordinary least squares; Pop, population; PRO, patient-reported outcome; Pts, patients; XGBC, eXtreme Gradient Boosting classification; XGBR, eXtreme Gradient Boosting regression.

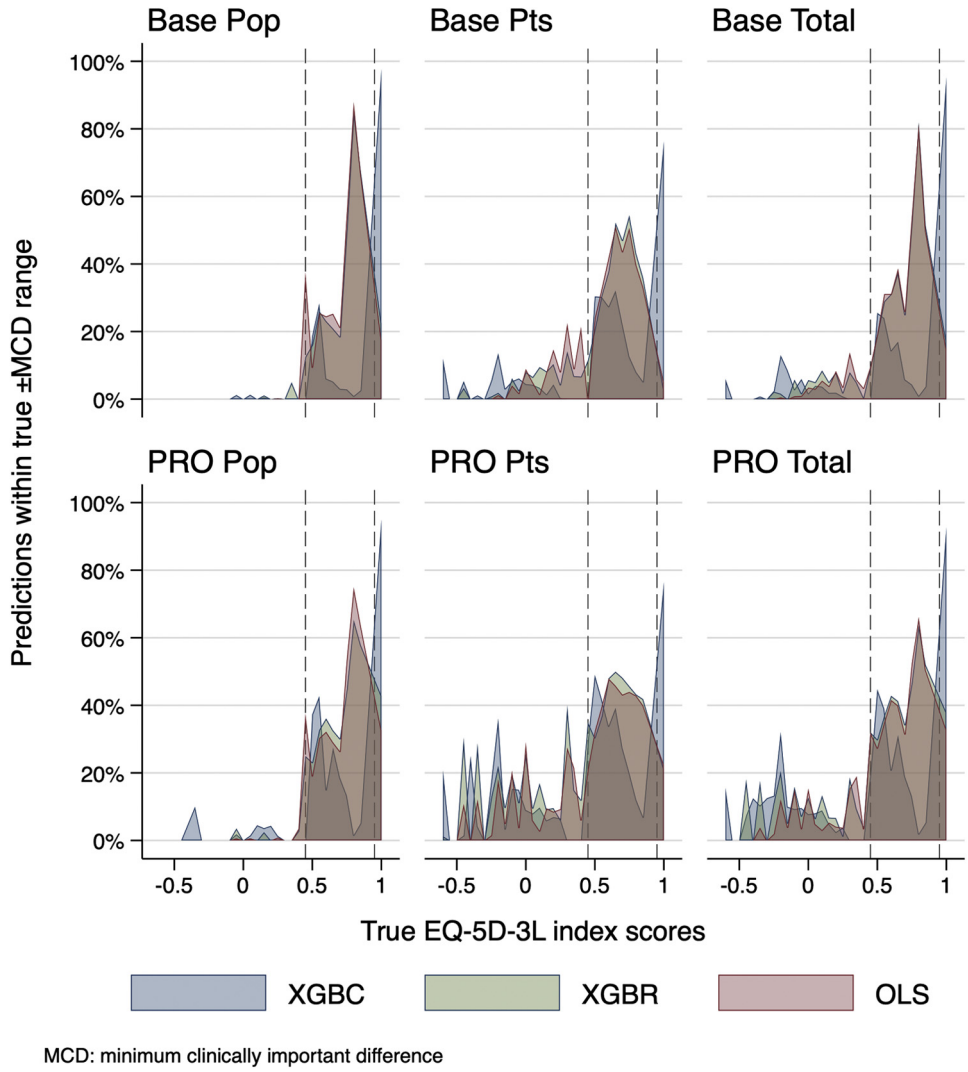
Figure 3. Percentage of predictions in the correct health severity group.



Error bars denote 95%CI of cross-validation
EQ-5D-3L index score ranges of health severity groups: low (< 0.45); medium (0.45 – 0.926); full health (> 0.926)

EQ-5D-3L indicates 3-level version of EQ-5D; OLS, ordinary least squares; Pop, population; PRO, patient-reported outcome; Pts, patients; XGBC, eXtreme Gradient Boosting classification; XGBR, eXtreme Gradient Boosting regression.

Figure 4. Percentage of predictions within clinically irrelevant error range by scenario and true EQ-5D-3L index scores.



EQ-5D-3L indicates 3-level version of EQ-5D; MCD, minimum clinically important difference; OLS, ordinary least squares; Pop, population; PRO, patient-reported outcome; Pts, patients; XGBC, eXtreme Gradient Boosting classification; XGBR, eXtreme Gradient Boosting regression.

patients (Appendix Fig. 4 of Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024). Regression to the mean was observed with all methods with positive bias (range 0.59-1.39) in low health and slight negative bias (range −0.01 to −0.23) in full health (Appendix Fig. 5 of Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024).

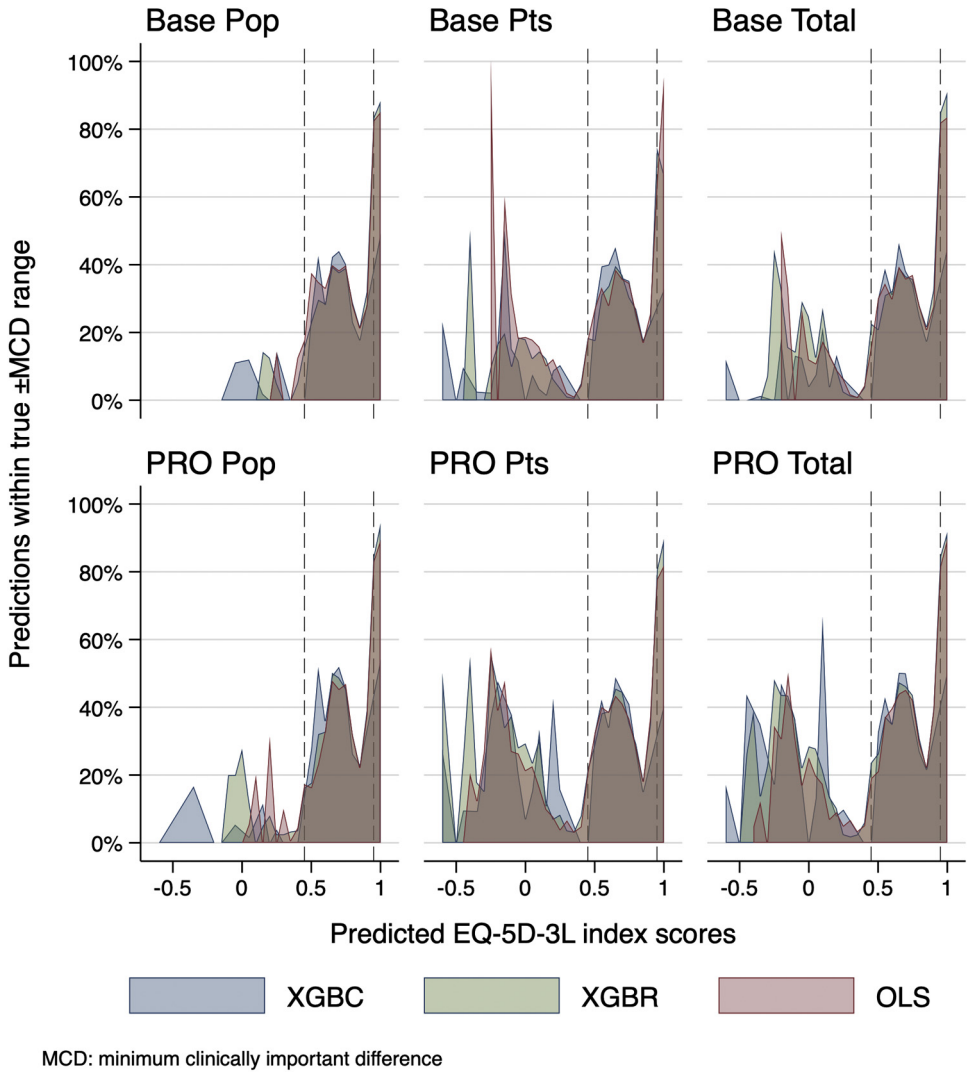
Accuracy of Predictions by True EQ-5D-3L Index Scores

Figure 4 illustrates the percent of predictions within clinically irrelevant error range by true EQ-5D-3L index scores. XGBC predictions were most accurate in the full health range, whereas XGBR and OLS predictions were most accurate in medium health. All methods were least accurate in the low health range, which improved after adding PROs to the base predictors. The proportions of predictions within the correct health severity category are depicted in Appendix Figure 6 of Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024.

Accuracy of Predictions by Predicted EQ-5D-3L Index Scores

If full health was predicted by XGBR or OLS, those values were mostly within, whereas full health predictions by XGBC were mostly outside the clinically irrelevant error range (Fig. 5). If medium health was predicted, the accuracy of the 3 methods was similar, albeit moderate. Low health predictions were the least accurate, which improved when PROs were added to base predictors, especially in patients. The pattern was similar for predictions in the correct health severity group. Although XGBC predictions were most accurate in the medium health range, the accuracy of XGBR and OLS improved from low health toward the full health range (Appendix Fig. 7 of Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024). Although mean prediction error of XGBR and OLS was approximately zero across the entire range, the bias of XGBC depended on the predicted EQ-5D-3L index values (Appendix Fig. 8 of Appendix 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.024).

Figure 5. Percentage of predictions within clinically irrelevant error range by scenario and predicted EQ-5D-3L index scores.



EQ-5D-3L indicates 3-level version of EQ-5D; MCD, minimum clinically important difference; OLS, ordinary least squares; Pop, population; PRO, patient-reported outcome; Pts, patients; XGBC, eXtreme Gradient Boosting classification; XGBR, eXtreme Gradient Boosting regression.

Discussion

We predicted EQ-5D-3L index scores via XGBC, XGBR, and OLS regression in a large international database combining multiple studies among patients and the general population with diverse predictors and a large amount of missing data. Across scenarios involving patients and the general population with and without PRO predictors, the percent of predictions within the clinically irrelevant error (true ± MCD) range were highest for XGBC and lowest for OLS with XGBR coming close. Nevertheless, MAE of prediction was lowest for XGBR followed by OLS and XGBC. Predictions with XGBC were biased. The performance of the 3 methods depended on the evaluation criteria, the target population, the predictor variables, and the EQ-5D-3L index range. Adding PROs to the demographic and disease-related predictors improved the accuracy of predictions.

Several studies have already used ML to predict EQ-5D values as a binary threshold[80,81] or as a continuous measure. Borchani et al[82] predicted EQ-5D-3L index scores from the 39-item Parkinson's Disease Questionnaire using multidimensional Bayesian network classifiers (MAE for OLS 0.350; MAE for multidimensional Bayesian network classifier 0.174). Gutacker et al[83] predicted postoperative health gains via classification and regression tree methodology (MAE ≤ root mean square error 0.158-0.224). Gao et al[84] mapped heart disease-specific quality of life to EQ-5D-5L index scores using econometric models and deep neural network algorithm (MAE for OLS 0.090-0.129; MAE for deep neural network 0.076-0.105). Recently, Mlynczak et al[85] applied random forest for assessing the construct validity of EQ-5D-5L (MAE ≤ root mean square error 0.121). In these studies, when compared, ML usually outperformed traditional econometric methods.[80,82,84] Advanced econometric models were also used to predict EQ-5D index values accommodating its multimodal distribution and upper limit at full health. In a rheumatoid arthritis data set, MAE was 0.1505 with linear regression, 0.1508 with a random effects Tobit model, 0.1508 with an adjusted limited

variable model (treating EQ-5D index score predictions > 0.883 as 1), and 0.1438 with a random effects adjusted limited variable mixture model.[86]

The strength of our study is that the analysis was performed on a large and diverse data set of multiple studies resembling real-world data connected in health data ecosystems.[14] MAE was comparable with previous studies using ML or regression methods. We evaluated prediction accuracy via the percent of predictions within the "clinically irrelevant error range" by splitting absolute error into "irrelevant" (≤MCD) and "relevant" (>MCD) values. We argue that there are no established criteria for further classifying errors into "large" or "acceptable" ones. Nevertheless, erroneous predictions in any direction of any magnitude may negatively affect decisions. Therefore, by conveying clinically relevant information about the shape of error distribution, this metric has merit in the evaluation of predictive models in healthcare.

Our study has limitations. Despite the full data matrices of individual studies, the joint database had a large proportion of missing data, which was handled via the missing indicator method. Although this method has been criticized for its biasedness,[87] it has recently been advocated in predictive or epidemiological research.[88,89] In our study, mean prediction error using XGBR and OLS was close to zero in all scenarios, whereas XGBC predictions were positively biased. Nevertheless, the performance of XGBoost is usually not affected by the imputation of missing data.[90] We did not apply multiple imputation techniques to prevent leakage of information and interference with the prediction methods. Therefore, the information contained in the data set probably could not be used to its full capacity. The potential effect of missing data structures on the predictive performance of the models deserves further exploration, along with the use of more advanced data imputation techniques such as multiple imputation[91] or LASSO regression.[92]

Although more advanced regression models are available to accommodate the multimodal distribution and upper limit of EQ-5D index scores at full health, simple OLS models are commonly applied to predict individual utilities.[86,93] Adding further interaction terms did not affect the performance of our OLS model. Nevertheless, variable selection via LASSO regression from many interacted predictors has been shown to improve the performance of OLS in predicting EQ-5D-5L index scores.[94] In addition, predictions were performed on unweighted data, which, through the nonrepresentative proportions of patients in the sample, may have introduced bias to the prediction results. Therefore, the external validity of our prediction models is probably limited and the accuracy of predictions may be further improved. As a future area of research, alternative prediction techniques, a combination of methods based on their performance in various EQ-5D-3L ranges, imputation of missing data, and weights reflecting the structure of the average population and disease epidemiology could be applied to improve the accuracy of predictions of individual EQ-5D-3L index scores. Furthermore, the external validity of the prediction models should be tested on multiple study populations that were not included in the training phase.

Conclusions

In a large database of EQ-5D-3L studies, prediction errors of EQ-5D-3L index scores using XGBC, XGBR, and OLS were greater than the MCD for most respondents and depended on the applied method, performance evaluation criteria, the target population, applied predictors, and the EQ-5D-3L range. The performance of XGBR slightly exceeded OLS in most evaluation measures. Regression methods outperformed XGBC in terms of prediction accuracy and bias.

Our results warn against overoptimistic expectations and prompt for care when using ML for predicting individual patient-reported health outcomes. We recommend the systematic and widespread collection of real-world PRO data using standardized PRO measures, including EQ-5D. In addition, we encourage dialogs between artificial intelligence and outcomes research experts to enhance the value of accumulating data in health systems.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.01.024.

Article and Author Information

Accepted for Publication: January 18, 2022

Published Online: March 15, 2022

doi: https://doi.org/10.1016/j.jval.2022.01.024

Author Affiliations: Health Economics Research Center, University Research and Innovation Center, Óbuda University, Budapest, Hungary (Zrubka, Gulácsi, Péntek); Corvinus Institue for Advanced Studies, Corvinus University of Budapest, Budapest, Hungary (Zrubka, Gulácsi); Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary (Csabai); Institute of Economics, Centre for Economic and Regional Studies, Budapest, Hungary (Hermann); Institute of Economics, Corvinus University of Budapest, Budapest, Hungary (Hermann); Department of Experimental and Clinical Pharmacology, Medical University of Warsaw, Warsaw, Poland (Golicki); Institute for Economic Research, Ljubljana, Slovenia (Prevolnik-Rupel, Ogorevc).

Correspondence: Zsombor Zrubka, MD, PhD, Health Economics Research Center; University Research and Innovation Center; Óbuda University, Bécsi út 96/b, Budapest 1034, Hungary. Email: zrubka.zsombor@uni-obuda.hu

Author Contributions: *Concept and design:* Zrubka, Csabai, Hermann, Golicki, Prevolnik-Rupel, Ogorevc, Gulácsi, Péntek
Acquisition of data: Golicki, Prevolnik-Rupel, Ogorevc, Gulácsi, Péntek
Analysis and interpretation of data: Zrubka, Csabai, Hermann
Drafting of the manuscript: Zrubka, Péntek
Critical revision of the paper for important intellectual content: Zrubka, Csabai, Hermann, Golicki, Prevolnik-Rupel, Ogorevc, Gulácsi, Péntek
Statistical analysis: Zrubka, Csabai, Hermann
Provision of study materials or patients: Golicki, Prevolnik-Rupel, Ogorevc, Gulácsi, Péntek
Obtaining funding: Zrubka, Gulácsi
Supervision: Gulácsi, Péntek

Conflict of Interest Disclosures: Drs Zrubka, Csabai, Hermann, Golicki, Prevolnik-Rupel, Gulácsi and Péntek reported receiving funding from project number TKP2020-NKA-02 implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme. Dr Zrubka reported receiving grants or contracts from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 679681); reported receiving consulting fees from Roche Hungary; and reported serving as a member of the ISPOR Digital Health Special Interest Group as co-chair of a key scientific project. Dr Csabai reported receiving funding from the Ministry of Innovation and Technology National Research Development and Innovation Office in the framework of the Artificial Intelligence National Laboratory Program (MILAB). Dr Golicki reported receiving research grants and travel reimbursement from the EuroQol Group. Dr Péntek reported receiving funding from project number 2019-1.3.1-KK-2019-00007 implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2019-1.3.1-KK funding scheme during writing this

study. Drs Golicki, Prevolnik-Rupel, and Péntek reported being members of the EuroQol Group. No other disclosures were reported.

Funding/Support: Under project number TKP2020-NKA-02, this research has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Berwick D, Black N, Cullen D, et al. Recommendations to OECD Ministers of Health from the high level reflection group on the future of health statistics: strengthening the international comparison of health system performance through patient-reported indicators. OECD. https://www.oecd.org/health/Recommendations-from-high-level-reflection-group-on-the-future-of-health-statistics.pdf. Accessed June 30, 2021.
2. Endpoints used for relative effectiveness assessment of pharmaceuticals. Health-related quality of life and utility measures. EUnetHTA. www.eunethta.eu. Accessed June 30, 2021.
3. Kennedy-Martin M, Slaap B, Herdman M, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *Eur J Health Econ*. 2020;21(8):1245–1257.
4. EQ-5D is a recommended tool for use in cost-utility analyses around the globe. EuroQoL Group. https://euroqol.org/eq-5d-is-a-recommended-tool-for-use-in-cost-utility-analyses-around-the-globe/. Accessed January 29, 2020.
5. EuroQoL Group. EuroQol — a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16(3):199–208.
6. EESZT National eHealth Infrastructure. National Centre for Healthcare Services. https://www.eeszt.gov.hu/hu/nyito-oldal. Accessed June 30, 2021.
7. Offerman A. Slovenia moving HIS applications to central eHealth platform (eZdravje). European Commission. Joinup Web Site. https://joinup.ec.europa.eu/collection/ehealth/document/slovenia-moving-his-applications-central-ehealth-platform-ezdravje. Accessed June 17, 2020.
8. Czerw A, Fronczak A, Witczak K, Juszczyk G. Implementation of electronic health records in Polish outpatient health care clinics — starting point, progress, problems, and forecasts. *Ann Agric Environ Med*. 2016;23(2):329–334.
9. Borges do Nascimento IJ, Marcolino MS, Abdulazeem HM, et al. Impact of big data analytics on people's health: overview of systematic reviews and recommendations for future studies. *J Med Internet Res*. 2021;23(4):e27275.
10. Boccia S, Pastorino R, Giraldi L, van den Bergen K. *Digitalisation and Big Data: Implications for the Health Sector*. Brussels, Belgium: European Parliament; 2018.
11. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94–98.
12. Garai A, Pentek I, Adamko A. Revolutionizing healthcare with IoT and cognitive, cloud-based telemedicine. *Acta Polytech Hung*. 2019;16(2):163–181.
13. Cohoon TJ, Bhavnani SP. Toward precision health: applying artificial intelligence analytics to digital health biometric datasets. *Per Med*. 2020;17(4):307–316.
14. Iacob N, Simonelli F. Towards a European health data ecosystem. *Eur J Risk Regul*. 2020;11(4):884–893.
15. Devlin NJ, Brooks R. EQ-5D and the EuroQol Group: past, present and future. *Appl Health Econ Health Policy*. 2017;15(2):127–137.
16. Ernstsson O, Janssen MF, Heintz E. Collection and use of EQ-5D for follow-up, decision-making, and quality improvement in health care — the case of the Swedish National Quality Registries. *J Patient Rep Outcomes*. 2020;4(1):78.
17. Rencz F, Gulacsi L, Drummond M, et al. EQ-5D in Central and Eastern Europe: 2000-2015. *Qual Life Res*. 2016;25(11):2693–2710.
18. Boros J, Györke J, Pásztorné Stokker E, Szabó ZK. *Results of the 2014 Health Interview Survey - summary data*. Budapest, Hungary: KSH; 2018.
19. 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.
20. Jia H, Lubetkin EI. Estimating EuroQol EQ-5D scores from Population Healthy Days data. *Med Decis Mak*. 2008;28(4):491–499.
21. Tsiachristas A, Potter CM, Rocks S, et al. Estimating EQ-5D utilities based on the Short-Form Long Term Conditions Questionnaire (LTCQ-8). *Health Qual Life Outcomes*. 2020;18(1):279.
22. Brodszky V, Balint P, Geher P, et al. Disease burden of psoriatic arthritis compared to rheumatoid arthritis, Hungarian experiment. *Rheumatol Int*. 2009;30(2):199–205.
23. Poor AK, Sardy M, Cserni T, et al. Assessment of health-related quality of life in psoriasis patients in Hungary. *Orv Hetil*. 2018;159(21):837–846.
24. Balogh O, Pentek M, Gulacsi L, et al. [Quality of life and burden of disease in peripheral arterial disease: a study among Hungarian patients]. *Orv Hetil*. 2013;154(12):464–470.
25. Pentek M, Kobelt G, Czirjak L, et al. Costs of rheumatoid arthritis in Hungary. *J Rheumatol*. 2007;34(6):1437.
26. Minier T, Pentek M, Brodszky V, et al. Cost-of-illness of patients with systemic sclerosis in a tertiary care centre. *Rheumatol (Oxf Engl)*. 2010;49(10):1920–1928.
27. Ersek K, Kovacs T, Wimo A, et al. Costs of dementia in Hungary. *J Nutr Health Aging*. 2010;14(8):633–639.
28. Brodszky V, Péntek M, Jelics N, et al. Health-related costs of diabetes mellitus in adults treated with insulin. Cross-sectional survey of 480 patients in general practice and outpatient settings. *Diabetologia Hungarica*. 2010;19(1):37–44.
29. Simoens S, Dunselman G, Dirksen C, et al. The burden of endometriosis: costs and quality of life of women with endometriosis and treated in referral centres. *Hum Reprod*. 2012;27(5):1292–1299.
30. Pentek M, Gulacsi L, Toth E, Baji P, Brodszky V, Horvath C. Ten-year fracture risk by FRAX((R)) of women with osteoporosis attending osteoporosis care in Hungary. *Orv Hetil*. 2016;157(4):146–153.
31. Pulay AJ, Bitter I, Papp S, et al. Exploring the relationship between quality of life (EQ-5D) and clinical measures in adult attention deficit hyperactivity disorder (ADHD). *Appl Res Qual Life*. 2016;12(2):409–424.
32. Hever NV, Pentek M, Ballo A, et al. Health related quality of life in patients with bladder cancer: a cross-sectional survey and validation study of the Hungarian version of the Bladder Cancer Index. *Pathol Oncol Res*. 2015;21(3):619–627.
33. Rencz F, Kovacs A, Brodszky V, et al. Cost of illness of medically treated benign prostatic hyperplasia in Hungary. *Int Urol Nephrol*. 2015;47(8):1241–1249.
34. Péntek M, Bereczki D, Gulácsi L, et al. Survey of epilepsy in adults in Hungary: quality of life and costs. *Ideggyógyászati Sz*. 2013;66(7–8):262.
35. Pentek M, Gulacsi L, Majoros A, et al. Health related quality of life and productivity of women with overactive bladder. *Orv Hetil*. 2012;153(27):1068–1076.
36. Tamas G, Gulacsi L, Bereczki D, et al. Quality of life and costs in Parkinson's disease: a cross sectional study in Hungary. *PLoS One*. 2014;9(9):e107704.
37. Péntek M, Harangozó J, Égerházi A, et al. Quality of life and disease burden of patients with schizophrenia in Hungary. *Psychiatr Hung*. 2012;27(1):4–17.
38. Pentek M, Gulacsi L, Rozsa C, et al. Health status and costs of ambulatory patients with multiple sclerosis in Hungary. *Ideggyogy Sz*. 2012;65(9-10):316–324.
39. Rencz F, Gulácsi L, Brodszky V, Golicki D, Ruzsa G, Péntek M. Pns401 the first parallel Eq-5d-3l and Eq-5d-5l composite time trade-off valuation study in Europe. *Value Health*. 2019;22.
40. Baji P, Brodszky V, Rencz F, Boncz I, Gulacsi L, Pentek M. Health status of the Hungarian population between 2000-2010. *Orv Hetil*. 2015;156(50):2035–2044.
41. Pentek M, Brodszky V, Biro Z, et al. Subjective health expectations of patients with age-related macular degeneration treated with antiVEGF drugs. *BMC Geriatr*. 2017;17(1):233.
42. Pentek M, Brodszky V, Gulacsi AL, et al. Subjective expectations regarding length and health-related quality of life in Hungary: results from an empirical investigation. *Health Expect*. 2014;17(5):696–709.
43. Donaldson C, Baker R, Mason H, et al. *European value of a quality adjusted life year — final publishable report*. Newcastle; 2010. https://www.researchgate.net/publication/327904832_European_Value_of_a_QALY_EuroVaQ_Final_Report_to_European_Commission. Accessed June 30, 2021.
44. Golicki D, Jakubczyk M, Niewada M, Wrona W, Busschbach JJ. Valuation of EQ-5D health states in Poland: first TTO-based social value set in Central and Eastern Europe. *Value Health*. 2010;13(2):289–297.
45. Golicki D, Niewada M. General population reference values for 3-level EQ-5D (EQ-5D-3L) questionnaire in Poland. *Pol Arch Med Wewn*. 2015;125(1-2):18–26.
46. Golicki D, Niewada M, Buczek J, et al. Validity of EQ-5D-5L in stroke. *Qual Life Res*. 2015;24(4):845–850.
47. Golicki D, Sliwka A, Fijewski G, Latek M. Pos14 quality of life according to EQ-5D after osteoporotic hip fracture in Poland. *Value Health*. 2006;9(6):A382–A383.
48. Golicki D, Zawodnik S, Janssen MF, Kiljan A, Hermanowski T. Eq1 psychometric comparison of Eq-5d and Eq-5d-5l in student population. *Value Health*. 2010;13(7):A240.
49. Prevolnik Rupel V, Srakar A, Rand K. Valuation of EQ-5D-3l Health States in Slovenia: VAS based and TTO based value sets. *Slovenian Journal of Public Health*. 2019;59(1):8–17.
50. Rupel VP, Rebojl M. *The Slovenian VAS tariff based on valuations of EQ-5D health states from the general population. EuroQol 17th Plenary Meeting*. Pamplona, Spain. 2000.
51. Rupel VP, Ogorevc M. Use of the EQ-5D instrument and value scale in comparing health states of patients in four health care programs among health care providers. *Value Health Reg Issues*. 2014;4:95–99.
52. Pentek M, Beretzky Z, Brodszky V, et al. Health-related productivity of the Hungarian population. A cross-sectional survey. *Orv Hetil*. 2020;161(36):1522–1533.
53. *Self-Reported Population Health: An International Perspective Based on EQ-5D*. Dordrecht, The Netherlands: Springer; 2014.
54. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997;35(11):1095–1108.

55. McCormack HM, Horne DJ, Sheather S. Clinical applications of visual analogue scales: a critical review. *Psychol Med*. 1988;18(4):1007–1019.
56. Folstein MF, Folstein SE, McHugh PR. Mini-mental state. *J Psychiatr Res*. 1975;12(3):189–198.
57. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*. 1983;33(11):1444–1452.
58. Fredriksson T, Pettersson U. Severe psoriasis—oral therapy with a new retinoid. *Dermatologica*. 1978;157(4):238–244.
59. Rencz F, Brodszky V, Gulacsi L, et al. Parallel valuation of the EQ-5D-3L and EQ-5D-5L by time trade-off in Hungary. *Value Health*. 2020;23(9):1235–1245.
60. International Statistical Classification of Diseases and Related Health Problems 10th Revision. https://icd.who.int/browse10/2016/ens. Accessed January 5, 2019.
61. Feng Y, Parkin D, Devlin NJ. Assessing the performance of the EQ-VAS in the NHS Proms programme. *Qual Life Res*. 2014;23(3):977–989.
62. Happiness in Nations: Overview of Happiness Surveys Using Measure Type. Erasmus University Rotterdam, Happiness Economics Research Organisation. http://worlddatabaseofhappiness.eur.nl. Accessed March 20, 2019.
63. Glossary: Minimum European Health Module (MEHM). Eurostat. https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Minimum_European_Health_Module_(MEHM). Accessed February 24, 2019.
64. Cox B, van Oyen H, Cambois E, et al. The reliability of the Minimum European Health Module. *Int J Public Health*. 2009;54(2):55–60.
65. Fries JF, Spitz PW, DY Y. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *J Rheumatol*. 1982;9(5):789–793.
66. Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI)—a simple practical measure for routine clinical use. *Clin Exp Dermatol*. 1994;19(3):210–216.
67. Mahoney F, Barthel D. Functional evaluation: the Barthel index. *Md State Med J*. 1965;14:61–65.
68. Zuckerman JD, Koval KJ, Aharonoff GB, Hiebert R, Skovron ML. A functional recovery score for elderly hip fracture patients: I. Development. *J Orthop Trauma*. 2000;14(1):20–25.
69. Gilbert SM, Dunn RL, Hollenbeck BK, et al. Development and validation of the Bladder Cancer Index: a comprehensive, disease specific measure of health related quality of life in patients with localized bladder cancer. *J Urol*. 2010;183(5):1764–1769.
70. Barry MJ, Fowler FJ, O'Leary MP, et al. The American Urological Association symptom index for benign prostatic hyperplasia. *J Urol*. 1992;148(5 Part 1):1549–1557.
71. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Paper presented at: KDD'16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; New York; 2016.
72. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev*. 2020;54(3):1937–1967.
73. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res*. 2005;30:79–82.
74. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res*. 2005;14(6):1523–1532.
75. Parkin D, Devlin N, Feng Y. What determines the shape of an EQ-5D index distribution? *Med Decis Mak*. 2016;36(8):941–951.
76. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ*. 2004;13(9):873–884.
77. Bramer M. *Principles of Data Mining*. London, United Kingdom: Springer; 2007.
78. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019;9(2):010318.
79. Tohka J, van Gils M. Evaluation of machine learning algorithms for health and wellness applications: a tutorial. *Comput Biol Med*. 2021;132:104324.
80. Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC Med Inform Decis Mak*. 2019;19(1):3.
81. Lee SK, Son YJ, Kim J, et al. Prediction model for health-related quality of life of elderly with chronic diseases using machine learning techniques. *Healthc Inform Res*. 2014;20(2):125–134.
82. Borchani H, Bielza C, Marti Nez-Marti NP, Larranaga P. Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: an application to predict the European Quality of Life-5 Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39). *J Biomed Inform*. 2012;45(6):1175–1184.
83. Gutacker N, Street A. Use of large-scale HRQoL datasets to generate individualised predictions and inform patients about the likely benefit of surgery. *Qual Life Res*. 2017;26(9):2497–2505.
84. 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.
85. Mlynczak K, Golicki D. Validity of the EQ-5D-5L questionnaire among the general population of Poland. *Qual Life Res*. 2021;30(3):817–829.
86. Hernandez Alava M, Wailoo AJ, Ara R. Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value Health*. 2012;15(3):550–561.
87. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ*. 2012;184(11):1265–1269.
88. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol*. 2020;125:183–187.
89. Song M, Zhou X, Pazaris M, Spiegelman D. The missing covariate indicator method is nearly valid almost always. Preprint. Posted online October 30, 2021. ArXiv:211100138. https://doi.org/10.48550/arXiv.2111.00138.
90. Rusdah DA, Murfi H. XGBoost in handling missing values for life insurance risk prediction. *SN Appl Sci*. 2020;2(8):1.
91. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
92. Lavanya K, Reddy L, Eswara Reddy B. Modeling of missing data imputation using additive lasso regression model in Microsoft Azure. *J Eng Appl Sci*. 2018;13:6324–6334.
93. 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.
94. Hay JW, Gong CL, Jiao X, et al. A US population health survey on the impact of COVID-19 using the EQ-5D-5L. *J Gen Intern Med*. 2021;36(5):1292–1301.