ScienceDirect

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

Preference-Based Assessments

Estimating Cost-Effectiveness Using Alternative Preference-Based Scores and Within-Trial Methods: Exploring the Dynamics of the Quality-Adjusted Life-Year Using the EQ-5D 5-Level Version and Recovering Quality of Life Utility Index

Matthew Franklin, BA, MSc, PhD, Rachael Maree Hunter, MSc, Angel Enrique, PhD, Jorge Palacios, MD, PhD, Derek Richards, PhD

ABSTRACT

Objectives: This study aimed to explore quality-adjusted life-year (QALY) and subsequent cost-effectiveness estimates based on the more physical health–focused EQ-5D 5-level version (EQ-5D-5L) value set for England or cross-walked EQ-5D 3-level version UK value set scores or more mental health recovery-focused Recovering Quality of Life Utility Index (ReQoL-UI), when using alternative within-trial statistical methods. We describe possible reasons for the different QALY estimates based on the interaction between item scores, health state profiles, preference-based scores, and mathematical and statistical methods chosen.

Methods: QALYs are calculated over 8 weeks from a case study 2:1 (intervention:control) randomized controlled trial in patients with anxiety or depression. Complete case and with missing cases imputed using multiple-imputation analyses are conducted, using unadjusted and regression baseline-adjusted QALYs. Cost-effectiveness is judged using incremental cost-effectiveness ratios and acceptability curves. We use previously established psychometric results to reflect on estimated QALYs.

Results: A total of 361 people (241:120) were randomized. EQ-5D-5L crosswalk produced higher incremental QALYs than the value set for England or ReQoL-UI, which produced similar unadjusted QALYs, but contrasting baseline-adjusted QALYs. Probability of cost-effectiveness <£30 000 per QALY ranged from 6% (complete case ReQoL-UI baseline-adjusted QALYs) to 64.3% (multiple-imputation EQ-5D-5L crosswalk unadjusted QALYs). The control arm improved more on average than the intervention arm on the ReQoL-UI, a result not mirrored on the EQ-5D-5L nor condition-specific (Patient-Health Questionnaire-9, depression; Generalized Anxiety Disorder-7, anxiety) measures.

Conclusions: ReQoL-UI produced contradictory cost-effectiveness results relative to the EQ-5D-5L. The EQ-5D-5L's better responsiveness and "anxiety/depression" and "usual activities" items drove the incremental QALY results. The ReQoL-UI's single physical health item and "personal recovery" construct may have influenced its lower 8-week incremental QALY estimates in this patient sample.

Keywords: anxiety, crosswalk, depression, economic evaluation, EQ-5D-5L, QALY, recovery, ReQoL-UI, preference-based.

VALUE HEALTH. 2021; ■(■):■–■

Introduction

Economic evaluation evidence helps inform resource allocation between alternative care interventions within a finite care budget.[1] Cost-effectiveness analysis via cost per quality-adjusted life-year (QALY) is recommended internationally, including by the National Institute for Health and Care Excellence (NICE) for England and Wales.[2–4] QALYs are measured on a preference-based quality-adjustment scale, anchored at 0 (a state equivalent to dead) and 1 (full health), combined with length of life allowing comparisons between interventions that affect quantity and quality of life.[1,5] Nevertheless, the concept of "a QALY is a QALY" for cross-comparable decision making has been debated extensively given that different preference-based measures and value sets produce different QALYs, stemming from aspects such as content and size of classification systems, and methods and populations used to value health states.[5–12] Additionally, alternative mathematical and statistical methods can influence QALY estimates and associated cost-effectiveness evidence.[13–15]

A more consistent, comparable approach is a rationale for NICE and reimbursement agencies internationally recommending the EQ-5D 3-level version (EQ-5D-3L) representing (3^5) 243 possible

health states as a generic health measure.[2-4] In comparison, the newer EQ-5D 5-level version (EQ-5D-5L) represents (5^5) 3125 possible health states resulting in increased sensitivity and reduced ceiling effects.[16-22] Country-specific EQ-5D-5L preference-based value sets are available (https://euroqol.org/), with the value set for England (VSE) based on a composite time trade-off (TTO) and discrete choice experiment hybrid model.[23-29] Nevertheless, an independent quality assurance study led to NICE recommending the van Hout et al crosswalk over the VSE.[30-34] Therefore, EQ-5D-5L preference-based values are cross-walked/mapped EQ-5D-3L UK value set scores based on the conventional TTO method.[35] Nevertheless, cross-walked scores have inherent concerns (eg, predictive errors) and do not represent a direct value set for the EQ-5D-5L.[36,37] Analyses internationally comparing EQ-5D-5L and EQ-5D-3L value sets and alternative cross-walked scores suggest that they estimate different preference-based values and subsequent QALYs.[38-41]

Related to mental health, the EQ-5D measures' underlying health domains/items (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) have been argued to be more physical than mental health focused, stimulating debate as to their appropriateness within mental health populations.[10,42-48] The 2010 Global Burden of Disease study estimated that depression and anxiety disorders contribute to a large portion of the total disability among all mental health and substance use disorders.[49] Approximately 1 in 6 adults in England has a common mental health disorder.[50] Mental health services and interventions have evolved to deal with care demand; for example, stepped-care within Improving Access to Psychological Therapies (IAPT) services in England and use of low-intensity interventions such as Digital Mental Health Interventions (DMHIs), which require appropriate cost-effectiveness evidence.[51-54] For reimbursement agencies such as NICE, alternative preference-based measures can be rationalized based on aspects such as psychometric performance (4; p. 42), as suggested by Brazier and Deverill.[55] EQ-5D measures' psychometric results offer better support in common (eg, anxiety and depression) than severe (eg, schizophrenia and bipolar disorder) mental health disorders.[44-47,56] The Recovering Quality of Life (ReQoL) 20-item (ReQoL-20) and ReQoL 10-item (ReQoL-10) versions are "recovery-focused quality of life" measures for mental health service users.[57] A UK value set using the composite TTO method has been developed to calculate QALYs from 7 ReQoL-10 items: the ReQoL-Utility Index (ReQoL-UI) representing (5^7) 78 125 possible health states.[58] The ReQoL-UI's developers suggest it is arguably a more mental health–focused generic measure relative to the more physical health–focused EQ-5D measures.[58] A psychometric analysis by Franklin and Enrique[59] in patients with anxiety and/or depression identified that, compared with the EQ-5D-5L using the VSE or UK crosswalk, the ReQoL-UI had better construct validity with depression severity, that is, Patient-Health Questionnaire-9 (PHQ-9) score,[60] whereby construct validity was assessed based on "convergent" (eg, correlation with the PHQ-9) and "known-group" validity (eg, assessing effect sizes between depression severity groupings; eg, "moderate" relative to "mild" severity). Nevertheless, the EQ-5D-5L preference-based score was more responsive (based on assessing standardized response means) and had better construct validity with anxiety severity, that is, Generalized Anxiety Disorder-7 (GAD-7) score.[59,61,62] These results suggest that the 2 preference-based measures may systematically differ in how they measure anxiety and depression, with implications for the precision of QALY estimation.[59]

We aim to explore the various QALY and subsequent cost-effectiveness estimates based on the EQ-5D-5L (VSE or crosswalk) or ReQoL-UI, when using alternative within-trial statistical methods based on a case study trial. Throughout we describe possible reasons for different QALY estimates based on the interaction between item scores, health state profiles, preference-based scores, and mathematical and statistical methods chosen, with suggested implications for evaluating interventions within mental health services such as IAPT and future research.

Methods

Data Source

A parallel-group, randomized waitlist-controlled trial examining the effectiveness and cost-effectiveness of internet-delivered cognitive behavioral therapy (iCBT) for patients presenting with depression or anxiety was conducted at an established IAPT service.[63,64] Before 2:1 randomization (intervention: 8-week waiting-list control), trial eligibility criteria were applied (see Appendix S1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358). Trial inclusion criteria were people (1) aged between 18 and 80 years, (2) above clinical thresholds for depression (PHQ-9 $\geq$10) or anxiety (GAD-7 $\geq$8),[60-62] and (3) suitable for iCBT (ie, willing to use iCBT, internet access). The structured Mini-International Neuropsychiatric Interview 7.0.2, administered by telephone by psychological wellbeing practitioners (ie, clinicians trained to deliver low-intensity support), established the presence or absence of a primary diagnosis of depression or anxiety disorder at baseline.[65] National Health Service England Research Ethics Committee provided trial ethics approval (Research Ethics Committee reference: 17/NW/0311). The trial was prospectively registered: current controlled trials ISRCTN91967124.

The trial is completed with the protocol and main results published showing that iCBT produced statistically significant improvements in depression (PHQ-9) and anxiety (GAD-7) severity compared with wait-list controls at 8 weeks, with further statistically significant intervention-group improvements from 8 weeks to 12 months.[63,64] Over 8 weeks, the probability of cost-effectiveness was 46.6% <£30 000 per EQ-5D-5L crosswalk-based QALY as the NICE reference case.[64] VSE and ReQoL-UI results were not published given NICE's VSE position and nonfinalized ReQoL-UI at point of submission.

Preference-Based Measures

The EQ-5D-5L is a self-reported, generic health measure with 5 severity levels over 5 dimensions/items.[22] VSE and crosswalk score range: −0.285 or −0.594 to 1, respectively.[25,34]

The ReQoL-UI classification system is based on 7 ReQoL-10 items: 3 positively (ReQoL-10 items: 5, 7, 10) and 3 negatively (ReQoL-10 items: 3, 6, 9) worded mental health items and its one physical health item. These 7 items cover 7 themes of self-reported recovery-focused quality of life[58]: autonomy, wellbeing, hope, activity, belonging and relationships, self-perception, and physical health. The ReQoL-UI is described as having 2 overall dimensions: a mental health (6 items) and a physical health (1 item) dimension.[58] ReQoL-UI score ranges from −0.195 to 1.[58]

Table 1. Preference-based score descriptive statistics for observed cases at baseline across and within-trial arms.

Parameter	Trial arm	n* (N %)	Mean	Median	SD	Min	Max	P. floor	P. ceiling	N floor (%)	N ceiling (%)	UHSP†	UPBS‡
EQ-5D-5L	Both	355 (98.3)	0.730	0.783	0.163	−0.010	1	−0.285	1	0 (0.0)	3 (0.8)	111	100
VSE	Int.	238 (98.8)	0.735	0.783	0.152	0.089	1	−0.285	1	0 (0.0)	2 (0.8)	83	75
	Cont.	117 (97.5)	0.722	0.783	0.182	−0.010	1	−0.285	1	0 (0.0)	1 (0.9)	57	54
EQ-5D-5L	Both	355 (98.3)	0.652	0.721	0.202	0.076	1	−0.594	1	0 (0.0)	3 (0.8)	111	105
crosswalk	Int.	238 (98.8)	0.656	0.718	0.193	0.119	1	−0.594	1	0 (0.0)	2 (0.8)	83	78
	Cont.	117 (97.5)	0.645	0.721	0.218	0.076	1	−0.594	1	0 (0.0)	1 (0.9)	57	56
ReQoL-UI	Both	353 (97.8)	0.778	0.807	0.141	0.115	0.995	−0.195	1	0 (0.0)	0 (0.0)	319	319
	Int.	237 (98.3)	0.788	0.806	0.123	0.242	0.979	−0.195	1	0 (0.0)	0 (0.0)	219	219
	Cont.	116 (96.7)	0.757	0.808	0.171	0.115	0.995	−0.195	1	0 (0.0)	0 (0.0)	114	114

Cont. indicates, control; EQ-5D-5L, EQ-5D 5-level version; Int., intervention; Max, maximum; Min, minimum; N, number of responders; P., possible; ReQoL-UI, Recovering Quality of Life-utility index; UHSP, unique health state profile; UPBS, unique preference-based score; VSE, value set for England.
*Number of participants at baseline was as follows: both trial arms ("Both"), N = 361; Int. arms, N = 241; Cont. arms, N = 120.
†UHSP: the descriptive system element of the EQ-5D-5L and ReQoL-UI questionnaires produces a 5-digit or 7-digit health state profile, respectively, that represents the level of reported problems on each of the 5 or 7 dimensions of health, for example, 11223 for the EQ-5D-5L or 1112234 for the ReQoL-UI. UHSP refers to the number of UHSPs represented by the group of participants on that specific measure, for example, across both trial arms (N = 361), 111 EQ-5D-5L health state profiles compared with 319 ReQoL-UI health state profiles are represented by the participant sample.
‡UPBS: UPBS refers to the number of UPBSs represented by the group. For the ReQoL-UI, UHSP and UPBS are equal such that each health state is represented by a UPBS. For the EQ-5D-5L VSE and crosswalk, the UPBS < UHSP as some health states are represented by the same preference-based score within the VSE or crosswalk; for example, on the VSE, a preference-based score of 0.469 represents 3 health state profiles (22235, 13415, 11434), and for the crosswalk, a preference-based score of 0.414 represents 2 health state profiles (12324, 11115) within our participant sample. NB our participant sample does not represent all possible UHSP and UPBS combinations; for example, using the crosswalk function, a preference-based score of 0.414 actually represents 6 health state profiles (11115, 12324, 14141, 24114, 31234, 41143) with the crosswalk producing 1079 UPBS relative to the 3125 UHSP the EQ-5D-5L purports to measure.

Economic Evaluation

This 8-week within-trial cost-effectiveness analysis focuses on the NICE reference case of cost-per-QALY from a health and social care perspective. Because estimated QALYs are the main interest here, intervention (£94.63 per person) and other cost calculations are described elsewhere.[64] We followed NICE guidelines, Consolidated Health Economic Evaluation Reporting Standards checklist, and recommended methods for handling preference-based (utility), cost, and missing data using Stata version 15 and Microsoft Excel 2016.[4,13-15,30,66-71]

Calculating QALYs

QALYs are calculated from preference-based scores using the total area under the curve (AUC) method[15]:

$$q_{jti} = \frac{\left(p_{j(t-1)i} + p_{jti}\right)}{2}\delta_t \quad (1)$$

whereby p, preference-based score; i, an individual; and t, time (ie, baseline, $t = 0$). For each group j ($j = 0$, control; $j = 1$, intervention), the consecutive time measures are added, averaged, and then rescaled (δ) for the percent of a year that t and t-1 cover, that is, 0.15 for 8 weeks. From Eq. (1), total QALYs (Q) for each individual's trial duration are the summation of QALY calculations for each follow-up time point starting at t = 1:

$$Q_{ji} = \sum_{t=1}^{T} q_{jti} \quad (2)$$

Preference-based scores at baseline (t = 0) and 8 weeks (t = 1) are reported alongside subsequent QALY estimates for both trial arms and from 8 weeks to 12 months (t = 5) for the intervention arm only.

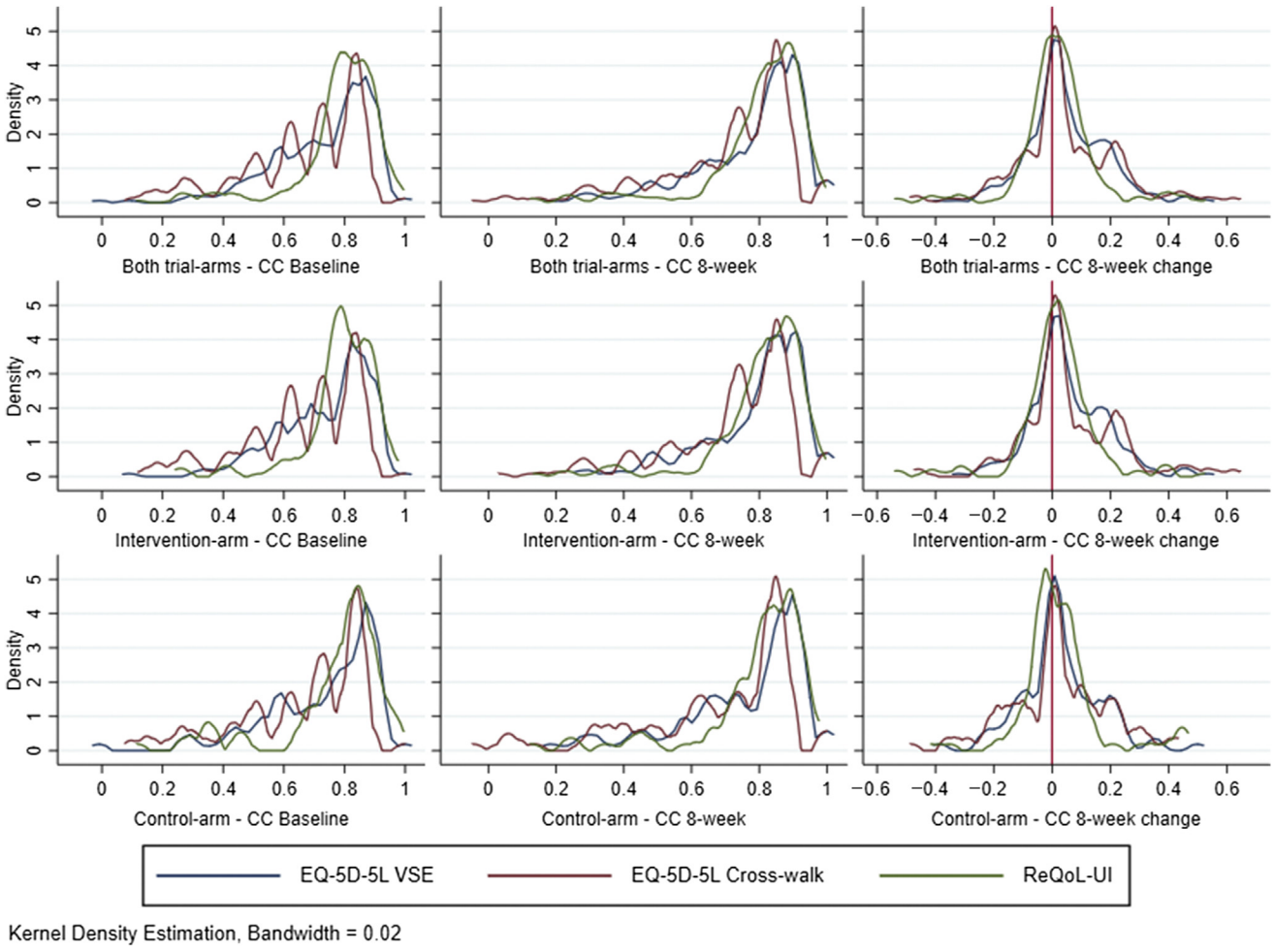
Statistical analyses

Analyses included complete cases (CCs) and with missing cases imputed based on multiple imputation (MI) by chained equation using predictive mean matching, drawing inference from a pool of 10 donors (k-nearest neighbors = 10) thus avoiding predicting missing values outside the plausible and observed range.[67,72] The MI method was chosen post hoc once the mechanism for missingness was deemed to be missing at random based on logistic regression which identified baseline sex, GAD-7 caseness, work and social adjustment scale, and IAPT phobia scale scores as predictors of missingness.[13-15,73] VSE, crosswalk, ReQoL-UI, and future cost missing cases at all follow-up time points were imputed. The number of imputed data sets was based on the percent of missing CC data across all time points in the intervention arm (m = 43).[13,74] Rubin's rule was applied when estimating MI analyses' means and standard errors of the mean (SEM).[75,76]

Baseline-adjusted QALYs are estimated using baseline preference-based values and trial arm as covariates within 2 independent regression models: ordinary least squares and seemingly unrelated regression, the latter accounting for the bivariate relationship between costs and QALYs.[15,77,78] Incremental mean-point estimates of trial arm differences (ie, intervention minus control) related to mean costs over mean QALYs are used to estimate incremental cost-effectiveness ratios (ICERs).

Bootstrapping was used to calculate bootstrapped 95% confidence intervals (bCIs) and bootstrapped SEMs around costs and QALYs and for plotting cost-effectiveness acceptability curves. CC and MI analyses involved 5000 or 21 500 (ie, 500 nested within imputed data sets: m = 43) bootstrapped iterations, respectively.[67] Cost-effectiveness acceptability curves present the probability of intervention cost-effectiveness compared with control across a range of cost-effectiveness thresholds, for example, NICE's £20 000 to £30 000 per QALY.[4] CC analyses bCIs are bias corrected and

Figure 1. CC smoothed EQ-5D-5L VSE and crosswalk, and ReQoL-UI score distributions across and within-trial arms—baseline, 8 weeks, and 8-week change.



CC indicates complete case; EQ-5D-5L, EQ-5D 5-level version; ReQoL-UI, ReQoL-Utility Index; VSE, value set for England.

accelerated 95% confidence intervals (95% BCa CIs), which corrects for the bias and skewness in the distribution of bootstrap estimates, which is methodologically complicated for MI data sets when jackknifing; therefore, percentile method bCIs (95% bCIs) are used to reflect value coverage across bootstrapped MI data sets.[13,76]

Additional analyses exploring the interaction between estimated QALYs, preference-based scoring algorithms, and item scores are described in the Appendices in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358.

Results

Descriptive Statistics

Overall, 361 people were randomized (241 intervention:120 control): 71.5% were female, with a mean age of 33 years (range 18-74). The Baseline Mini-International Neuropsychiatric Interview 7.0.2 diagnosis is as follows: 52%, major depressive disorder; 64%, anxiety disorder; and 36%, both. The CC and MI analysis included 282 (194:88) and 352 participants (236:116), respectively. Appendices S1-S3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358 includes a Consolidated Standards of Reporting Trials diagram, further demographic details, and measure completeness statistics.

Preference-Based Scores

Table 1 provides preference-based score descriptive statistics for observed cases at baseline across and within-trial arms. The crosswalk suggests this patient sample has the lowest, and the ReQoL-UI suggests the highest, mean preference-based health status at baseline. The EQ-5D-5L suggests this patient population is less heterogeneous than the ReQoL-UI, categorizing 355 participants into 111 unique health state profiles (UHSPs), whereas the ReQoL-UI categorizes 353 participants into 319 UHSPs. Relatedly, each ReQoL-UI UHSP is accompanied by its own unique preference-based score (UPBS). In comparison, 111 UHSPs are quantified by 100 VSE UPBSs and 105 crosswalk UPBSs, because some health states are represented by the same preference-based score (see Table 1).

Figure 1 shows kernel density estimates for the CC analyses preference-based scores at baseline and 8 weeks, as plotted on a graph within and across trial arms; the change in score over this 8-week period is also presented. Figure 1 shows the VSE's

Table 2. Summary of costs, baseline and 8-week scores, unadjusted and BA QALYs by measure in the CC and MI analysis over the 8-week trial period.

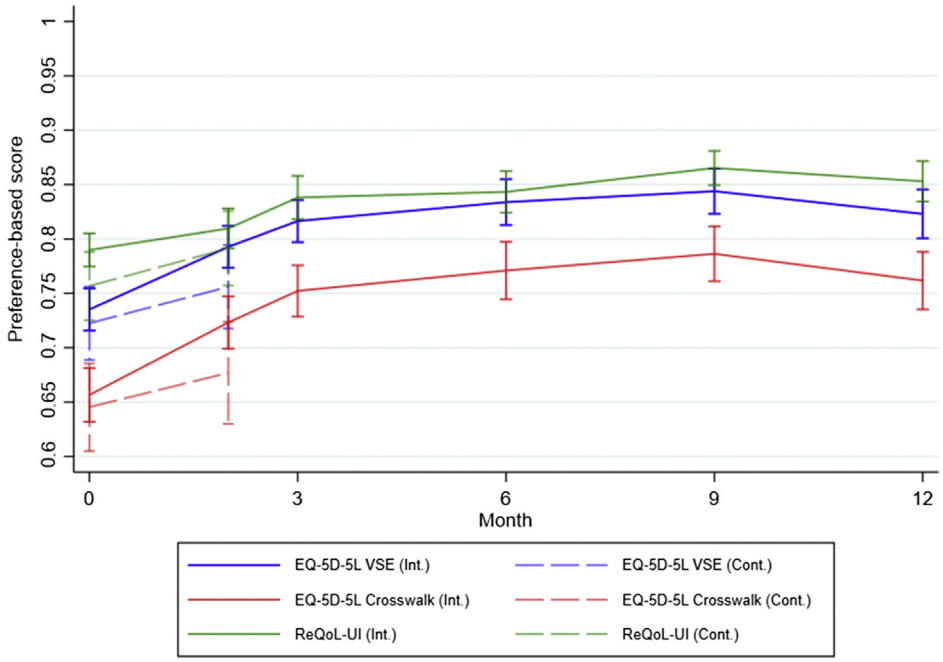
Measure	Metric	Intervention					Control					Dif. trial arm	
		Mean	SEM*	95% CI†		Dif., t_1-t_0 [BA]‡	Mean	SEM*	95% bCI†		Dif. t_1-t_0 [BA]‡	Mean	Dif. t_1-t_0 [BA]‡
CC—8 weeks		n = 194 (80%)					n = 88 (73%)						
Costs	£, t_1	£197.29	£18.22	£170.55	£247.34	—	£110.76	£35.54	£69.48	£245.72	—	£86.53	—
EQ-5D-5L	Score, t_0	0.7362	0.0108	0.7150	0.7574	—	0.7278	0.0197	0.6893	0.7663	—	0.0084	—
VSE	Score, t_1	0.7931	0.0106	0.7723	0.8138	0.0568	0.7571	0.0196	0.7188	0.7954	0.0293	0.0360	0.0275
	QALYs	0.1176	0.0015	0.1145	0.1203	—	0.1142	0.0028	0.1083	0.1193	—	0.0034	—
	BA QALYs	0.1173	0.0006	0.1129	0.1170	−0.0003	0.1150	0.0011	0.1160	0.1186	0.0007	0.0024	−0.0011
EQ-5D-5L	Score, t_0	0.6587	0.0136	0.6320	0.6853	—	0.6564	0.0224	0.6125	0.7004	—	0.0022	—
crosswalk	Score, t_1	0.7226	0.0132	0.6968	0.7485	0.0639	0.6767	0.0250	0.6278	0.7257	0.0203	0.0459	0.0436
	QALYs	0.1063	0.0018	0.1024	0.1095	—	0.1026	0.0033	0.0954	0.1086	—	0.0037	—
	BA QALYs	0.1062	0.0008	0.1045	0.1078	−0.0001]	0.1027	0.0014	0.1000	0.1055	[0.0002]	0.0034	−0.0003
ReQoL-UI	Score, t_0	0.7921	0.0086	0.7753	0.8090	—	0.7663	0.0192	0.7287	0.8039	—	0.0258	—
	Score, t_1	0.8086	0.0101	0.7888	0.8284	0.0165	0.7915	0.0176	0.7569	0.8260	0.0251	0.0171	−0.0087
	QALYs	0.1231	0.0012	0.1204	0.1253	—	0.1198	0.0026	0.1140	0.1242	—	0.0033	—
	BA QALYs	0.1222	0.0007	0.1208	0.1235	−0.0010	0.1220	0.0009	0.1201	0.1239	0.0022	0.0001	−0.0032
MI—8 weeks		N = 236 (98%)					N = 116 (97%)						
Costs	£, t_1	£196.86	£18.06	£166.19	£235.92	—	£109.93	£35.04	£64.06	£193.79	—	£86.94	—
EQ-5D-5L	Score, t_0	0.7352	0.0098	0.7159	0.7545	—	0.7225	0.0170	0.6892	0.7558	—	0.0127	—
VSE	Score, t_1	0.7928	0.0097	0.7738	0.8119	0.0576	0.7560	0.0193	0.7181	0.7939	0.0335	0.0368	0.0241
	QALYs	0.1175	0.0013	0.1148	0.1202	—	0.1137	0.0025	0.1088	0.1184	—	0.0038	—
	BA QALYs	0.1170	0.0002	0.1167	0.1175	−0.0005	0.1147	0.0007	0.1137	0.1161	0.0010	0.0023	−0.0015
EQ-5D-5L	Score, t_0	0.6566	0.0125	0.6320	0.6812	—	0.6453	0.0204	0.6054	0.6852	—	0.0112	—
crosswalk	Score, t_1	0.7232	0.0122	0.6993	0.7471	0.0666	0.6769	0.0236	0.6306	0.7232	0.0315	0.0463	0.0351
	QALYs	0.1061	0.0017	0.1027	0.1094	—	0.1017	0.0029	0.0958	0.1075	—	0.0044	—
	BA QALYs	0.1057	0.0003	0.1053	0.1062	−0.0004	0.1026	0.0008	0.1012	0.1040	0.0009	0.0031	−0.0013
ReQoL-UI	Score, t_0	0.7899	0.0077	0.7748	0.8051	—	0.7567	0.0159	0.7257	0.7878	—	0.0332	—
	Score, t_1	0.8096	0.0093	0.7914	0.8278	0.0197	0.7914	0.0172	0.7577	0.8250	0.0346	0.0183	−0.0149
	QALYs	0.1230	0.0011	0.1208	0.1250	—	0.1191	0.0022	0.1146	0.1231	—	0.0040	—
	BA QALYs	0.1218	0.0002	0.1213	0.1220	−0.0013	0.1217	0.0006	0.1205	0.1228	0.0026	0.0001	−0.0039

Note. CC sample size based on responders at baseline and 8 weeks: 198 intervention arms (82%) and 91 control arms (76%); MI sample size across all time points: 236 intervention arms (98%) and 116 control arms (97%).
BA indicates baseline adjusted; BCa, bias corrected and accelerated; bCI, bootstrapped confidence intervals; CC, complete cases analysis; CI, confidence interval; Dif. mean, difference in mean values between trial arms; EQ-5D-5L, EQ-5D 5-level version; ICER, incremental cost-effectiveness ratio; MI, multiple imputation; MICE, multiple imputation using chained equations; QALY, quality-adjusted life-year; ReQoL-UI, ReQoL-Utility Index; SEM, standard error of the mean; VSE, value set for England.
*Preference-based scores: SEMs are calculated using Rubin's rule for the MI analysis; (BA) QALYs: SEMs are bootstrapped.
†In the CC analysis, bCIs are BCa; in the MI analysis, bCI are based on the Percentile method.
‡Score (Dif., t_1-t_0), the difference in mean score at 8 weeks (t_1) minus mean score at baseline (t_0); QALYs (Dif. [BA]), the difference in mean BA QALY minus mean (unadjusted) QALY.

distribution is "smoother" than for the crosswalk, but not the ReQoL-UI, which is partly due to the number of UHSPs and UPBSs represented by each measure. Smoother in this context implies a broader distribution of scores across the score range resulting in less clustering and lower density around specific score ranges dependent on the prespecified bandwidth (ie, 0.02 for Fig. 1). Nevertheless, particularly at baseline, the ReQoL-UI presents higher density at the upper end of the scale (eg, >0.7) than the crosswalk or VSE, which can relatively restrict ability for greater ReQoL-UI improvement after baseline. Relatedly in the intervention arm, the ReQoL-UI's high central density just above zero for 8-week score change is similar to the VSE and crosswalk, but the VSE and crosswalk have a broader distribution and additional peaks (eg, >0.15), which contributes to a greater mean change.

Figure 2 presents MI mean and 95% confidence interval preference-based scores across all data collection time points and up to 8 weeks in Table 2. These results suggest crosswalk-based health is poorer than that estimated using the VSE or ReQoL-UI, which are more similar with each other than the crosswalk (Fig. 2). The ReQoL-UI suggests that over 8 weeks the mean difference in preference-based health between trial arms decreases, whereas the EQ-5D-5L suggested it increased with implications for estimating incremental QALYs. In the intervention arm, a statistically significant difference with baseline preference-based scores is achieved by 8 weeks for the EQ-5D-5L but not until 3

Figure 2. MI preference-based score means with 95% CIs at baseline and 8 weeks per trial arm and at 3, 6, 9, and 12 months in the intervention arm.



CI indicates confidence interval; Cont. indicates control; EQ-5D-5L, EQ-5D 5-level version; Int., intervention; MI, multiple imputation; ReQoL-UI, ReQoL-Utility Index; VSE, value set for England.

months for the ReQoL-UI; this 3-month period represents the natural treatment timeframe in the intervention arm not captured by the 8-week comparative trial period nor the incremental QALY estimates.

Incremental Results

Table 2 indicates the crosswalk produces the largest incremental QALY difference between trial arms over 8 weeks, although the ReQoL-UI produces more incremental QALYs than the VSE suggesting the opposite to the change in preference-based scores (Fig. 2 and Table 2). This is because baseline imbalances are not accounted for across the individuals' total AUC calculations (Eq. 1), with regression-based adjustment recommended over individual-level adjustment as part of the AUC calculation.[78,79] Regression-based baseline-adjustment using the total AUC takes into account baseline imbalances in preference-based scores and the phenomenon that those individuals with preference-based scores that are lower or higher than the mean at baseline will usually experience a respectively higher or lower improvement at follow-up. Therefore, because of the baseline imbalance and greater variation between the 2 arms in the ReQoL-UI, when a baseline-adjustment is statistically applied, the mean incremental difference in QALYs between the 2 arms is smaller than without the baseline-adjustment.[78]

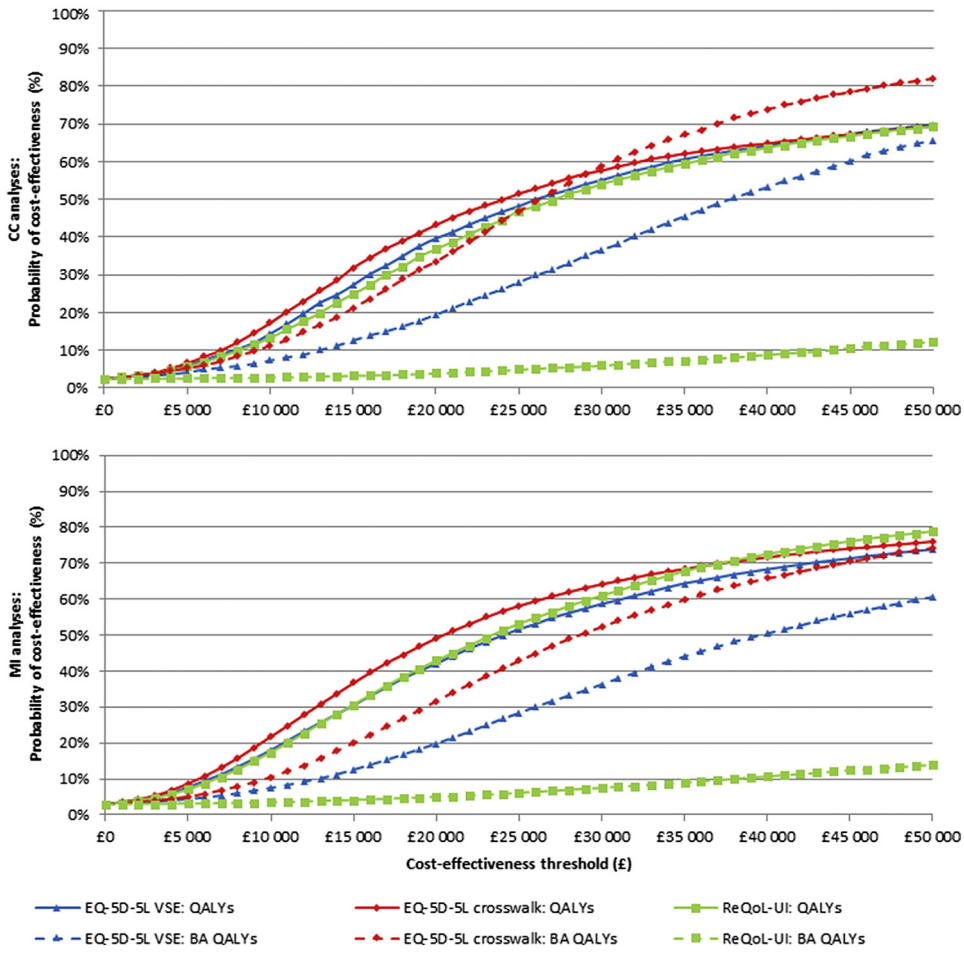
Table 3 and Figure 3 show that across both CC and MI unadjusted analyses, EQ-5D-5L and ReQoL-UI suggest iCBT is cost-effective <£30 000 per QALY (probability range 54%-64%). Baseline-adjusted QALY results are contrary to the aforementioned, whereby for the same MI analyses the ReQoL-UI suggests the highest ICER (£1 252 542) relative to the crosswalk's lowest "cost-effective" ICER (£27 684). When accounting for baseline-adjusted QALYs across CC and MI analyses, probability of cost-effectiveness <£30 000 per QALY ranged from 6% (CC ReQoL-UI baseline-adjusted QALY) to 58.9% (CC crosswalk baseline-adjusted QALY). The largest change in probability of cost-effectiveness when moving from unadjusted to baseline-adjusted QALYs was for the ReQoL-UI in the MI analysis, which dropped from 60.9% to 7.4%—an absolute decrease of 53.5%. Baseline-adjusted costs and seemingly unrelated regression results are presented in Appendix S4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358.

The change in EQ-5D-5L and ReQoL-UI item-level scores are described in Appendix S5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358. To summarize, the EQ-5D-5L's cost-effectiveness results seem to be driven by the "usual activities" and "anxiety/depression" items, with the intervention arm having better outcomes on average than the control arm across all EQ-5D-5L domains. Nevertheless, ReQoL-UI's item results were more varied, with the control group having better outcomes on average than the intervention arm across 3 (belonging and relationship, physical activity, self-perception) of its 7 items, influencing the incremental ReQoL-UI results and subsequent QALY estimates.

Discussion

This study supports current empirical evidence that value sets such as the VSE and cross-walked scores produce different QALYs even when from the same classification system.[38,39,41,80] We found that the VSE preference-based scores were more similar to those from the ReQoL-UI than the crosswalk. This meant the VSE and ReQoL-UI produced similar unadjusted QALYs. These similarities disappeared when statistically accounting for baseline preference-based scores, given that the control group improved more on average than the intervention group over 8 weeks on the ReQoL-UI—a result not mirrored on the EQ-5D-5L nor the trial's

Figure 3. CEACs for the CC (top) and MI (bottom) analyses dependent on BA and unadjusted QALYs.



BA indicates baseline adjusted; CC, complete case; CEAC, cost-effectiveness acceptability curve; EQ-5D-5L, EQ-5D 5-level version; MI, multiple imputation; QALY, qualityadjusted life-year; ReQoL-UI, ReQoL-Utility Index; VSE, value set for England.

condition-specific (GAD-7 and PHQ-9) measures.[64] This meant the ReQoL-UI had a lower probability of the intervention being cost-effective than the VSE or crosswalk: a decision maker is unlikely to consider implementing iCBT based on these ReQoL-UI results, but might when using the crosswalk results. These differences stem from the analyses conducted (eg, CC vs MI; unadjusted vs baseline adjusted) and the measures themselves.

Exploring Why the ReQoL-UI and EQ-5D-5L Produce Different QALYs

The different preference-based scores produced by the ReQoL-UI and EQ-5D-5L stem from aspects such as the content and size of their classification systems, the methods and populations used to value health states, and how their underlying preference-based scoring algorithms are constructed.

The ReQoL-UI can quantify a larger number of health states than the EQ-5D-5L (ie, 78 125 vs 3125), suggesting our study sample are more heterogeneous by categorizing them into almost 3 times more health state profiles than the EQ-5D-5L. This categorization stems from responses at the item-score level, which indicated more response variability for the ReQoL-UI than the EQ-5D-5L (see Appendix S5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358). The ability to categorize population samples into more health states should permit the measure to be more sensitive to change in generic health status, as long as that change is represented by the measure's items and preference-based score.

As far as the current authors are aware, there is only one published psychometric assessment of the EQ-5D-5L and ReQoL-UI—a study conducted by the current authors using the same data source as this article specifically to inform the associated within-trial economic evaluation.[59] This psychometric analysis suggests the ReQoL-UI has poorer responsiveness to change in GAD-7 anxiety or PHQ-9 depression severity than the EQ-5D-5L, which will have contributed to the smaller incremental QALY gains observed in this within-trial economic evaluation. Additionally, although the EQ-5D-5L was identified as having better construct validity with GAD-7 anxiety severity than the ReQoL-UI, the ReQoL-UI had better construct validity with PHQ-9 depression severity. The items that drove these construct validity results, particularly for the EQ-5D-5L, were the same items for which we identified a statistically significant difference between trial arms over 8 weeks (eg, "anxiety/depression" and "usual activities") as shown in Appendix S5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358.[59]

The ReQoL-UI has some perceived benefits over the EQ-5D-5L in mental health populations, including the ability to represent a larger number and variety of mental health states with better depression construct validity. Nevertheless, in the MI analysis, for example, the

Table 3. Incremental cost-effectiveness results by preference-based measure score in the CC and MI analysis over the 8-week trial period.

Measure	Metric	Trial arm differences					Mean
		Mean	bSEM	95% bCI[†]		Dif. BA	ICER, £
CC—8 weeks							
Costs	£, t_1	£86.53	£40.32	–£44.48	£140.39	-	-
EQ-5D-5L	QALY	0.0034	0.0031	–0.0020	0.0101	-	25 348
VSE	*BA QALY*	0.0024	0.0012	0.0000	0.0049	–0.0011	36 803
EQ-5D-5L	QALY	0.0037	0.0038	–0.0030	0.0117	-	23 385
crosswalk	*BA QALY*	0.0034	0.0016	0.0004	0.0067	–0.0003	25 287
ReQoL-UI	QALYs	0.0033	0.0028	–0.0017	0.0096		26 192
	BA QALY	<0.0001	0.0012	–0.0021	0.0025	–0.0032	577 331
MI—8 weeks							
Costs	£, t_1	£86.94	£38.47	–£2.26	£150.32	-	-
EQ-5D-5L	QALY	0.0038	0.0027	–0.0015	0.0094	-	22 828
VSE	*BA QALY*	0.0023	0.0014	–0.0004	0.0051	–0.0015	37 561
EQ-5D-5L	QALY	0.0044	0.0033	–0.0022	0.0112	-	19 624
crosswalk	*BA QALY*	0.0031	0.0018	–0.0003	0.0066	–0.0013	27 684
ReQoL-UI	QALY	0.0040	0.0024	–0.0006	0.0089	-	21 966
	BA QALY	<0.0001	0.0012	–0.0023	0.0027	–0.0039	1 252 542

Note. The CC analysis is based on 194 people (of 241; 80%) in the intervention arm and 88 (of 120; 73%) in the control arm; the MI analysis is based on 236 people (of 241; 98%) in the intervention arm and 116 (of 120; 97%) in the control arm.
BA indicates baseline adjusted; BCa, bias corrected and accelerated; bCI, bootstrapped confidence interval; bSEM: bootstrapped standard error of the mean; CC, complete cases analysis; CE, cost-effectiveness; Dif. mean, difference in mean values between trial arms; E, East; EQ-5D-5L, EQ-5D 5-level version; ICER, incremental cost-effectiveness ratio; k, thousand; MI, multiple imputation; Prob. CE, probability of cost-effectiveness; NE, North East; NW, North West; Q, QALYs; QALY: quality-adjusted life-year; ReQoL-UI, ReQoL-Utility Index; SE, South East; SW, South West; VSE, value set for England.
[*]CE plane quadrants are SE (less costly, more effective), SW (less costly, less effective), NE (more costly, more effective), NW (more costly, less effective), and E (more effective).
[†]In the CC analysis, bCIs are BCa; in the MI analysis, bCI are based on the percentile method.

continued on next page

incremental ReQoL-UI baseline-adjusted QALYs were minimal (<0.0001) compared with those estimated from the VSE (0.0023) or crosswalk (0.0031), a result stemming in part from the ReQoL-UI's poorer responsiveness (particularly over 8 weeks). This is an unexpected result given that we would expect the ReQoL-UI to be more responsive given its mental health focus and classification system.

Nevertheless, the psychometric analysis only partly explains the different QALY estimations. Also influencing the result is that the ReQoL-UI's preference-based score is based on a "random effects model consisting of a quadratic specification of ɵ (newtheta) with interaction terms for ɵ and levels 3, 4, and 5 of physical health[sic]."[58] In other words, the physical health item or dimension has a direct interaction with the mental health dimension (ɵ) within the ReQoL-UI preference-based scoring algorithm. This is practically and conceptually different to how the EQ-5D value sets are scored with implications for the derived preference-based score. It is important that researchers currently using, or considering using, the ReQoL-UI are aware of this interaction and associated rationale as described by Keetharuth and Rowen.[58] It is our hypothesis that the interaction with the physical health item contributed to the responsive statistics identified by the previous psychometric analysis and why the control group improved more on average than the intervention group in this IAPT-based within-trial analysis, as discussed further in the next sub-section and Appendix S6 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358.[59]

Implications for Mental Health Services, Users, and Research

The trial context is important for interpreting our results. IAPT step 2 focuses on specific mental health populations and interventions; that is, common mental health conditions that could benefit from low-intensity therapies as brief psychological interventions (eg, DMHI, Bibliotherapy) offered with support from clinicians.[81] Furthermore, IAPT standards of patient recovery focus on symptom improvement, where "recovery" is defined as moving from "caseness" (PHQ-9 ≥10; GAD-7 ≥8) to "no caseness."[54]

The ReQoL-UI psychometrics and within-trial results are potentially representative of its intended "recovery-focused" construct, which is different to "recovery" as operationalized by IAPT. Such symptomatic changes seem to be captured in part by the EQ-5D-5L dimensions of "usual activities" and "anxiety/depression," which drive our within-trial results (Appendix S5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358). In comparison, the ReQoL-UI is developed from a conceptual framework of personal recovery in mental health, which is more focused on improving long-term wellbeing through self-management and having personally meaningful life goals, therefore expanding beyond the traditional symptom-based recovery paradigm.[57,82-85] Given that IAPT performance metrics are, in part, symptom-based recovery with a focus on mental health, previous psychometric results suggest that the EQ-5D-5L captures these aspects better for anxiety severity and with greater

Table 3. Continued

Measure	Metric	ICERs by CE plane quadrant (%)*						Prob. CE < CE threshold (%)			
		SE: >Q| <£, %	SW: <Q| <£, %	NE: >Q| >£, %	NW: <Q| >£, %	E: >Q, %	Dif. BA, %	<£20k, %	Dif. BA, %	<£30k, %	Dif. BA, %
CC—8 weeks											
Costs	£, t_1	-	-	-	-	-	-	-	-	-	-
EQ-5D-5L	QALY	2.1	0.4	84.7	12.8	86.8	-	39.6	-	55.1	-
VSE	*BA QALY*	2.5	0.1	94.9	2.6	97.4	10.6	19.5	−20.1	36.7	−18.4
EQ-5D-5L	QALY	1.9	0.6	82.4	15.1	84.4	-	43.3	-	57.7	-
crosswalk	*BA QALY*	2.5	0.1	95.9	1.5	98.4	14.0	33.6	−9.7	58.9	1.2
ReQoL-UI	QALYs	2.2	0.3	85.9	11.6	88.1	-	36.8	-	54.0	-
	BA QALY	1.2	1.3	52.5	44.9	53.8	−34.3	4.0	−32.9	6.0	−48.0
MI—8 weeks											
Costs	£, t_1	-	-	-	-	-	-	-	-	-	-
EQ-5D-5L	QALY	2.7	0.0	88.4	8.9	91.0	-	42.1	-	58.6	-
VSE	*BA QALY*	2.6	0.1	92.3	5.0	94.9	3.9	19.8	−22.3	36.3	−22.3
EQ-5D-5L	QALY	2.6	0.1	86.9	10.4	89.5	-	49.2	-	64.3	-
crosswalk	*BA QALY*	2.6	0.1	93.7	3.6	96.3	6.8	31.7	−17.5	52.3	−11.9
ReQoL-UI	QALY	2.7	0.0	92.7	4.6	95.3	-	42.9	-	60.9	-
	BA QALY	1.7	1.0	48.3	49.0	50.1	−45.3	4.9	−37.9	7.4	−53.5

responsiveness than the ReQoL-UI, and this is reflected in our IAPT-based within-trial economic evaluation results.[59]

Additionally, as mentioned in the previous sub-section, the ReQoL-UI's preference-based scoring algorithm includes a physical health interaction term with the mental health domain; this type of interaction term is not used in the EQ-5D measures' preference-based value set scoring algorithms. Step 2 IAPT patients are referred on the basis of experiencing acute depression and/or anxiety symptomology, with improvements in physical health not being a key purpose of the service. In this trial's IAPT-based population, the majority of participants reported baseline physical health as "no problem" or "slight problem," with the majority not moving from this baseline state (Appendix S5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358). The interaction term in the ReQoL-UI's preference-based scoring algorithm means that because the majority of the study sample have no or slight problems with baseline physical health from which there is no change over the trial period, there is subsequently restricted ability for the ReQoL-UI's preference-based score to change, even if there are changes across the mental health domain. This will have influenced the ReQoL-UI's responsiveness, but also incremental QALY estimates, particularly given that the control arm randomly had more people who reported worse physical health at baseline and had a higher mean improvement in physical health over 8 weeks than the intervention arm. Compared with the ReQoL-UI, for the EQ-5D value sets, an interaction term is not imposed between the physical and mental health items allowing more independence between items in the preference-based scoring algorithm—this aspect is explored further in Appendix S6 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1358.

In different mental health settings (eg, hospital outpatients) and patient populations (severe mental health disorders), with different intervention types (high-intensity interventions), these psychometric and within-trial results could be different. Further research is warranted including to what extent various mental health interventions, from medication to DMHIs, are intended to promote symptomatic or personal recovery and physical health, which itself could dictate whether the EQ-5D-5L or ReQoL-UI may be the more appropriate preference-based measure to estimate cost-effectiveness. Further exploratory analysis of the ReQoL-UI is warranted before it is used to guide resource-allocation decision making, particularly as a complement or substitute to the EQ-5D-5L. Additionally, EuroQol's blog provides updates for its new health and wellbeing instrument (EQ-HWB), which should be considered for future research.[86]

Limitations

The 8-week between trial arm analyses limited the ability to assess incremental QALYs over a longer time-horizon. Common mental health disorder trials rarely exceed 12-month follow-up, with most follow-up periods aligning with when clinical change is most likely to be observed following treatment: between 8 and 12 weeks.[87-89] The lack of longer-term data also limits the ability and/or reliability of conducting extrapolated or modeling-based analyses over an even longer time-horizon. A systematic review of DMHI economic evaluations stated that 54 of 66 included articles did not explore the results beyond trial endpoints: "lack of longer-term modeling is likely to be due to, in part, the lack of reliable data about the long-term performance of DMHIs."[51] These data-driven limitations suggest longer-term comparative trial follow-ups are needed whenever possible with statistical methods as secondary options.[14,90,91]

The VSE has suggested complications beyond what our analysis explores, with a new UK valuation study underway.[31-33,59,92] Nevertheless, as an imperfect direct value set for the EQ-5D-5L relative to the crosswalk that represents the EQ-5D-3L UK value set, it is still useful and informative for this exploratory analysis.

Conclusions

These results indicate the importance of reflecting on a preference-based measure's whole design before using it for economic evaluation, aspects of which can be revealed by conducting psychometric analyses, given that on QALY face value it is difficult to wholly understand why different preference-based measures produce different QALYs. These differences stem from mathematical and statistical methods used and the preference-based measure itself, which need to be considered holistically to understand any subsequent QALY and cost-effectiveness estimates before suggesting to decision makers if an intervention is "cost-effective" or not based on such evidence.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2021.11.1358.

Article and Author Information

Accepted for Publication: November 1, 2021

Published Online: xxxx

doi: https://doi.org/10.1016/j.jval.2021.11.1358

Author Affiliations: Health Economics and Decision Science (HEDS), School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, England, UK (Franklin); Research Department of Primary Care and Population Health, Royal Free Medical School, University College London, London, England, UK (Hunter); Clinical Research & Innovation, SilverCloud Health, Dublin, Ireland (Enrique, Palacios, Richards); E-mental Health Research Group, School of Psychology, Trinity College, University of Dublin, Dublin, Ireland (Enrique, Palacios, Richards).

Correspondence: Matthew Franklin, BA, MSc, PhD, Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, West Court, 1 Mappin St, Sheffield, England, United Kingdom S1 4DT. Email: matt.franklin@sheffield.ac.uk

Author Contributions: *Concept and design:* Franklin, Enrique, Palacios, Richards
Acquisition of data: Enrique, Richards
Analysis and interpretation of data: Franklin, Hunter, Enrique, Palacios, Richards
Drafting of the manuscript: Franklin, Hunter, Enrique, Palacios, Richards
Critical revision of the paper for important intellectual content: Franklin, Hunter, Enrique, Palacios, Richards
Statistical analysis: Franklin
Obtaining funding: Franklin, Enrique, Palacios, Richards

Conflict of Interest Disclosures: Dr Franklin reported other from SilverCloud Health and other from National Institute for Health Research Applied Research Collaboration Yorkshire and Humber, during the conduct of the study. Drs Enrique, Palacios, and Richards are employees of SilverCloud Health. No other disclosures were reported. The views expressed are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Funding/Support: The trial from which the data for analysis were obtained was funded by SilverCloud Health. Study resources for the trial from Berkshire Healthcare National Health Services Foundation Trust, including research and development support, psychological wellbeing practitioners, case managers, and lead clinicians have been generously given in kind for the purpose of trial execution. The analysis and writing of the manuscript was partly funded by the National Institute for Health Research Applied Research Collaboration Yorkshire and Humber (National Institute for Health Research award identifier: 200166).

Role of the Funder/Sponsor: Employees of SilverCloud Health had a role in the design and conduct of the study; collection, management, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The National Institute for Health Research 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. The funding agreement ensured the authors' independence in developing the purview of the manuscript, writing, and publishing the manuscript.

Acknowledgment: The authors thank the research and development and clinical team members at Berkshire Healthcare National Health Service Foundation Trust service for assisting trial execution: Judith Chapman, Gabriella Clark, Emma Cole, and Sarah Sollese. The authors thank our colleagues at SilverCloud Health for providing administrative support and assisting data collection. The authors thank Anju Keetharuth, Donna Rowen, and John Brazier at ScHARR, University of Sheffield, for answering our questions in regard to the Recovering Quality of Life-Utility Index. The authors also thank the many patients who volunteered their time and efforts to participate in the trial.

REFERENCES

1. Drummond MF, Sculpher MJ, Claxton K, et al. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford, United Kingdom: Oxford University Press; 2015.
2. Rowen D, Azzabi Zouraq I, Chevrou-Severac H, van Hout B. International regulations and recommendations for utility data for health technology assessment. *Pharmacoeconomics*. 2017;35(suppl 1):11–19.
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. Guide to the methods of technology appraisal 2013. National Institute for Health and Care Excellence (NICE). https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781. Accessed March 10, 2021.
5. Brazier J, Ratcliffe J, Saloman J, et al. *Measuring and Valuing Health Benefits for Economic Evaluation*. Oxford, United Kingdom: Oxford University Press; 2016.
6. Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. *Pharmacoeconomics*. 2017;35(suppl 1):21–31.
7. Brazier J, Rowen D. NICE DSU technical support document 11: alternatives to EQ-5D for generating health state utility values. National Institute for Health and Care Excellence (NICE). https://www.ncbi.nlm.nih.gov/books/NBK425861/pdf/Bookshelf_NBK425861.pdf. Accessed March 10, 2021.
8. Rowen D, Brazier J, Ara R, Azzabi Zouraq I. The role of condition-specific preference-based measures in health technology assessment. *Pharmacoeconomics*. 2017;35(suppl 1):33–41.
9. Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA. Condition-specific preference-based measures: benefit or burden? *Value Health*. 2012;15(3):504–513.
10. Brazier J. Measuring and valuing mental health for use in economic evaluation. *J Health Serv Res Policy*. 2008;13(suppl 3):70–75.
11. Lancsar E, Gu Y, Gyrd-Hansen D, et al. The relative value of different QALY types. *J Health Econ*. 2020;70:102303.
12. Weinstein MC. A QALY is a QALY–or is it? *J Health Econ*. 1988;7(3):289–290.
13. Faria R, Gomes M, Epstein D, White IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *Pharmacoeconomics*. 2014;32(12):1157–1170.
14. Franklin M, Lomas J, Walker S, Young T. An educational review about using cost data for the purpose of cost-effectiveness analysis. *Pharmacoeconomics*. 2019;37(5):631–643.
15. Hunter RM, Baio G, Butt T, Morris S, Round J, Freemantle N. An educational review of the statistical issues in analysing utility data for cost-utility analysis. *Pharmacoeconomics*. 2015;33(4):355–366.
16. Golicki D, Niewada M, Karlińska A, et al. Comparing responsiveness of the EQ-5D-5L, EQ-5D-3L and EQ VAS in stroke patients. *Qual Life Res*. 2015;24(6):1555–1563.
17. Buchholz I, Thielker K, Feng YS, Kupatz P, Kohlmann T. Measuring changes in health over time using the EQ-5D 3L and 5L: a head-to-head comparison of measurement properties and sensitivity to change in a German inpatient rehabilitation sample. *Qual Life Res*. 2015;24(4):829–835.

18. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. *Value Health*. 2008;11(2):275–284.
19. Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Med Care*. 2007;45(3):259–263.
20. Scalone L, Ciampichini R, Fagiuoli S, et al. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. *Qual Life Res*. 2013;22(7):1707–1716.
21. Golicki D, Zawodnik S, Janssen MF, et al. Psychometric comparison of EQ-5D and EQ-5D-5L in student population. *Value Health*. 2010;13. A240-A240.
22. 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.
23. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*. 2014;17(4):445–453.
24. Stolk E, Ludwig K, Rand K, van Hout B, Ramos-Goñi JM. Overview, update, and lessons learned from the International EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. *Value Health*. 2019;22(1):23–30.
25. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. *Health Econ*. 2018;27(1):7–22.
26. Devlin NJ, Tsuchiya A, Buckingham K, Tilling C. A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. *Health Econ*. 2011;20(3):348–361.
27. Janssen BM, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. *Eur J Health Econ*. 2013;14(suppl 1):S5–S13.
28. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. *Med Care*. 2017;55(7):e51–e58.
29. Rowen D, Brazier J, Van Hout B. A comparison of methods for converting DCE values onto the full health-dead QALY scale. *Med Decis Making*. 2015;35(3):328–340.
30. Position statement on use of the EQ-5D-5L valuation set for England. National Institute for Health and Care Excellence (NICE). https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l. Accessed March 10, 2021.
31. Hernández-Alava M, Pudney S, Wailoo A. *Quality Review of a Proposed EQ-5D-5L Value Set for England* [EEPRU report] [online]; 2018.
32. Norman R, Olsen JA. Competing views on the English EQ-5D-5L valuation set. *Value Health*. 2020;23(5):574–575.
33. van Hout B, Mulhern B, Feng Y, Shah K, Devlin N. The EQ-5D-5L value set for England: response to the "Quality Assurance". *Value Health*. 2020;23(5):649–655.
34. van Hout B, Janssen M, Feng YS, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health*. 2012;15(5):708–715.
35. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997;35(11):1095–1108.
36. Mukuria C, Rowen D, Harnan S, et al. An updated systematic review of studies mapping (or Cross-Walking) measures of health-related quality of life to generic preference-based measures to generate utility values. *Appl Health Econ Health Policy*. 2019;17(3):295–313.
37. Longworth L, Rowen D. NICE DSU technical support document 10: The use of mapping methods to estimate health state utility values. National Institute for Health and Care Excellence (NICE). https://www.ncbi.nlm.nih.gov/books/NBK425834/pdf/Bookshelf_NBK425834.pdf. Accessed March 10, 2021.
38. Gerlinger C, Bamber L, Leverkus F, et al. Comparing the EQ-5D-5L utility index based on value sets of different countries: impact on the interpretation of clinical study results. *BMC Res Notes*. 2019;12(1):18.
39. Mulhern B, Feng Y, Shah K, et al. Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets [published correction appears in *Pharmacoeconomics*. 2018;36(6):727]. *Pharmacoeconomics*. 2018;36(6):699–713.
40. Hernandez Alava M, Wailoo A, Grimm S, et al. EQ-5D-5L versus EQ-5D-3L: the impact on cost effectiveness in the United Kingdom. *Value Health*. 2018;21(1):49–56.
41. Wailoo A, Alava MH, Pudney S, et al. An international comparison of EQ-5D-5L and EQ-5D-3L for use in cost-effectiveness analysis. *Value Health*. 2021;24(4):568–574.
42. Franklin M. Cost utility analysis. In: Razzouk D, ed. *Mental Health Economics*. Switzerland: Springer; 2017.
43. Razzouk D. *Mental Health Economics: The Costs and Benefits of Psychiatric Care*. Switzerland: Springer; 2017.
44. Brazier J, Connell J, Papaioannou D, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health Technol Assess*. 2014;18(34). vii–188.
45. Mulhern B, Mukuria C, Barkham M, et al. Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D. *Br J Psychiatry*. 2014;205(3):236–243.
46. Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. *Pharmacoeconomics*. 2015;33(11):1137–1154.
47. Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *Eur J Health Econ*. 2018;19(4):557–570.
48. Longworth L, Yang Y, Young T, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health Technol Assess*. 2014;18(9):1–224.
49. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet*. 2013;382(9904):1575–1586.
50. McManus S, Bebbington P, Jenkins R, Brugha T. Mental health and wellbeing in England: adult psychiatric morbidity survey 2014. NHS Digital. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/556596/apms-2014-full-rpt.pdf. Accessed March 10, 2021.
51. Jankovic D, Bojke L, Marshall D, et al. Systematic review and critique of methods for economic evaluation of digital mental health interventions. *Appl Health Econ Health Policy*. 2021;19(1):17–27.
52. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *Int Rev Psychiatry*. 2011;23(4):318–327.
53. Gyani A, Shafran R, Layard R, Clark DM. Enhancing recovery rates: lessons from year one of IAPT. *Behav Res Ther*. 2013;51(9):597–606.
54. The Improving Access to Psychological Therapies Manual. NHS Digital. https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/. Accessed December 12, 2021.
55. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Econ*. 1999;8(1):41–51.
56. 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.
57. Keetharuth AD, Brazier J, Connell J, et al. Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. *Br J Psychiatry*. 2018;212(1):42–49.
58. Keetharuth AD, Rowen D, Bjorner JB, Brazier J. Estimating a Preference-Based Index for mental health from the Recovering Quality of Life measure: valuation of Recovering Quality of Life Utility Index. *Value Health*. 2021;24(2):281–290.
59. Franklin M, Enrique A, Palacios J, Richards D. Psychometric assessment of EQ-5D-5L and ReQoL measures in patients with anxiety and depression: construct validity and responsiveness. *Qual Life Res*. 2021;30(9):2633–2647.
60. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606–613.
61. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Ann Intern Med*. 2007;146(5):317–325.
62. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. 2006;166(10):1092–1097.
63. Richards D, Duffy D, Blackburn B, et al. Digital IAPT: the effectiveness & cost-effectiveness of internet-delivered interventions for depression and anxiety disorders in the Improving Access to Psychological Therapies programme: study protocol for a randomised control trial. *BMC Psychiatry*. 2018;18(1):59.
64. Richards D, Enrique A, Eilert N, et al. A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. *npj Digit Med*. 2020;3:1–10.
65. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry*. 1998;59(suppl 20):22–57.
66. Microsoft excel 2016. Microsoft Corporation.
67. Leurent B, Gomes M, Faria R, Morris S, Grieve R, Carpenter JR. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial [published correction appears in Pharmacoeconomics. 2019;37(7):971]. *Pharmacoeconomics*. 2018;36(8):889–901.
68. StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC; 2017.
69. Ramsey S, Willke R, Briggs A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. *Value Health*. 2005;8(5):521–533.
70. Ramsey SD, Willke RJ, Glick H, et al. Cost-effectiveness analysis alongside clinical trials II—an ISPOR Good Research Practices Task Force report. *Value Health*. 2015;18(2):161–172.
71. Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS) – explanation and elaboration: a report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. *Value Health*. 2013;16(2):231–250.
72. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol*. 2014;14:75.
73. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2002.
74. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377–399.
75. Little RJ, Rubin DB. *Statistical Analysis With Missing Data*. Hoboken, NJ: John Wiley & Sons; 2019.

76. Burton A, Billingham LJ, Bryan S. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. *Clin Trials*. 2007;4(2):154–161.
77. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Econ*. 2004;13(5):461–475.
78. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Econ*. 2005;14(5):487–496.
79. Richardson G, Manca A. Calculation of quality adjusted life years in the published literature: a review of methodology and transparency. *Health Econ*. 2004;13(12):1203–1210.
80. Hernández Alava M, Pudney S, Wailoo A. Estimating the Relationship Between EQ-5D-5L and EQ-5D-3L: Results From an English Population Study [EEPRU Report]. University of Sheffield & University of York. http://www.eepru.org.uk/wp-content/uploads/2020/10/eq5d-5l-final-report-30-9-20.pdf. Accessed March 10, 2021.
81. Bennett-Levy J, Farrand P, Christensen H, et al. *Oxford Guide to Low Intensity CBT Interventions*. Oxford, United Kingdom: Oxford University Press; 2010.
82. Leamy M, Bird V, Le Boutillier C, Williams J, Slade M. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. *Br J Psychiatry*. 2011;199(6):445–452.
83. Shepherd G, Boardman J, Rinaldi M, et al. Supporting recovery in mental health services: quality and outcomes. Implementing Recovery Organ Change (ImROC). https://imroc.org/resources/8-supporting-recovery-mental-health-services-quality-outcomes/. Accessed March 10, 2021.
84. Slade M, Longden E. Empirical evidence about recovery and mental health. *BMC Psychiatry*. 2015;15:285.
85. Onken SJ, Craig CM, Ridgway P, Ralph RO, Cook JA. An analysis of the definitions and elements of recovery: a review of the literature. *Psychiatr Rehabil J*. 2007;31(1):9–22.
86. EuroQol is developing a new instrument - the EQ-HWB. EuroQol. https://euroqol.org/euroqol-is-developing-a-new-instrument-the-eq-hwb-2/. Accessed March 10, 2021.
87. Ramsberg J, Asseburg C, Henriksson M. Effectiveness and cost-effectiveness of antidepressants in primary care: a multiple treatment comparison meta-analysis and cost-effectiveness model. *PLoS One*. 2012;7(8):e42003.
88. Annemans L, Brignone M, Druais S, De Pauw A, Gauthier A, Demyttenaere K. Cost-effectiveness analysis of pharmaceutical treatment options in the first-line management of major depressive disorder in Belgium. *Pharmacoeconomics*. 2014;32(5):479–493.
89. Depression in adults: recognition and management. Clinical guidance [CG90]. National Institute for Health and Care Excellence (NICE). https://www.nice.org.uk/guidance/cg90. Accessed March 10, 2021.
90. Richardson J, Khan MA, Iezzi A, Maxwell A. Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D multiattribute utility instruments. *Med Decis Making*. 2015;35(3):276–291.
91. Briggs A, Sculpher M, Claxton K. *Decision Modelling for Health Economic Evaluation*. Oxford, United Kingdom: Oxford University Press; 2006.
92. New UK EQ-5D-5L valuation study. EuroQol. https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/valuation-standard-value-sets/new-uk-eq-5d-5l-valuation-study_blog/. Accessed March 10, 2021.