medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Predicting Quality Adjusted Life Years in young people attending primary mental health services

Matthew P Hamilton[†,1,*] Caroline X Gao[†,1,2,3] Kate M Filia[1,2]
Jana M Menssink[1,2] Sonia Sharmin[1,2] Nic Telford[4] Helen Herrman[1,2]
Ian B Hickie[5] Cathrine Mihalopoulos[6] Debra J Rickwood[7,4]
Patrick D McGorry[1,2] Sue M Cotton[1,2]

Abstract

Background: Quality Adjusted Life Years (QALYs) are often used in economic evaluations, yet utility weights for deriving them are rarely directly measured in mental health services.

Objectives: We aimed to: (i) identify the best Transfer To Utility (TTU) algorithms and predictors for an adolescent specific Multi-Attribute Utility Instrument - the Assessment of Quality of Life - six dimensions (AQoL-6D) and (ii) assess ability of TTU algorithms to predict longitudinal change.

Methods: We recruited 1107 young people attending Australian primary mental health services, collecting data at two time points, three months apart. Five linear and three generalised linear models were explored to identify the best TTU algorithm. Forest models were used to explore predictive ability of six candidate measures of psychological distress, depression and anxiety and linear / generalised linear mixed effect models were used to construct longitudinal predictive models for AQoL-6D change.

Results: A depression measure (Patient Health Questionnaire-9) was the strongest independent predictor of health utility. Linear regression models with complementary log-log transformation of utility score were the best preforming models. Between-person associations were slightly larger than within-person associations for most of the predictors.

Conclusions: Adolescent AQoL-6D utility can be derived from a range of psychological distress, depression and anxiety measures. TTU algorithms estimated from cross-sectional data may slightly bias QALY predictions.

Toolkits: The TTU models produced by this study can be searched, retrieved and applied to new data to generate QALY predictions with the Youth Outcomes to Health Utility (youthu) R package - https://ready4-dev.github.io/youthu.

[†] These authors contributed equally to this work.

[1] Orygen, Parkville, Australia
[2] Centre for Youth Mental Health; University of Melbourne, Parkville, Australia
[3] School of Public Health and Preventive Medicine, Monash University, Clayton, Australia
[4] headspace National Youth Mental Health Foundation, Melbourne, Australia
[5] Brain and Mind Centre, Youth Mental Health & Technology, Faculty of Medicine & Health, The University of Sydney, Australia
[6] Deakin Health Economics, Institute for Health Transformation, Deakin University, Geelong, Australia
[7] Faculty of Health, University of Canberra, Australia

[*] Correspondence: Matthew P Hamilton <matthew.hamilton@orygen.org.au>

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1 Introduction

To efficiently allocate scarce public resources between competing mental health programs, it is useful to have a common measure of benefit, ideally with a broadly accepted economic value. Quality adjusted life years (QALYs) are generic indices of outcome that inform public health policy in many countries [1] and are frequently used in economic evaluations across a range of health areas, including mental health. The "quality" in QALYs is often measured via the use of health related quality of life measures (commonly called multi-attribute utility instruments (MAUIs)), where the "importance" of the domains of quality of life measured by the questionnaire are weighted using the preferences of people [2]. This scoring approach produces a single health utility weight for each individual for each measured health state that is anchored on a 0 to 1 scale, where 0 represents death and 1 represents perfect health. Health utility weighs can be converted to QALYs by weighting the duration (the "years" part of QALYs) each individual spends in each health state.

MAUIs are regularly collected in research studies such as clinical trials and epidemiological surveys, but rarely in administrative health care records and treatment evaluation datasets. In the absence of direct measurement, Transfer to Utility (TTU) analysis has been developed to map utility weights from standard health status measurements [3]. In mental health settings, TTU algorithms have been developed to map psychological distress (measured using Kessler Psychological Distress Scale – 10 items, K10) and depression and anxiety symptoms (measured using Depression, Anxiety, and Stress Scale – 21 items, DASS-21 [4]) to a range of health utility measures including the Assessment of Quality of Life – 8 dimensions (AQoL-8D [5]). Published mental health TTU algorithms have been developed for adult [5] or child [6] general populations; however, they have questionable appropriateness for predicting health utility in clinical mental health samples of young people. Other difficulties with currently available TTU algorithms include over-reliance on cross-sectional data (not capturing the longitudinal dimension of QALYs), and on a limited range of predictors that are not routinely collected in youth mental health services.

With a sample of help-seeking young people attending primary mental health care services, we aimed to: (i) identify the best TTU regression models to predict AQoL-6D utility and evaluate the predictive ability of six candidate measures of psychological distress (Kessler Psychological Distress Scale - 6 Item (K6)), depression (Patient Health Questionnaire (PHQ-9) and Behavioural Activation for Depression Scale (BADS)) and anxiety (Generalised Anxiety Disorder Scale (GAD-7), Screen for Child Anxiety Related Disorders (SCARED) and Overall Anxiety Severity and Impairment Scale (OASIS)); and (ii) assess ability of the TTU algorithms to predict longitudinal (three-month) change.

2 Methods

2.1 Sample and setting

This study forms part of a larger research program focused on developing better outcome measures for young people seeking mental health support, and the study sample has previously been described [7]. Briefly, young people aged 12 to 25 years who presented for a first appointment for mental health or substance use related issues were recruited from three metropolitan and two regional Australian youth-focused primary mental health clinics (*headspace* centres) between September 2016 to April 2018. Sample characteristics are similar to previous descriptions of headspace clients, with slight differences in age (less 12-14 year olds, more 18-20 year olds), cultural background (more Culturally and Linguistically Diverse and less Aboriginal and Torres Strait Islander young people), sexuality (fewer heterosexual clients) and housing (more in unstable accommodation) [7].

2.2 Measures

We collected data on utility weights, six candidate predictors of utility weights including psychological distress, depression and anxiety measures as well as demographic, clinical and functional population information.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.2.1 Utility weights

We assessed utility weights using the adolescent version of the Assessment of Quality of Life – Six Dimension scale (AQoL-6D; [8]) MAUI. It was selected due to the relevance of its domains for a clinical mental health sample [9] and its acceptable participant time-burden. The adolescent AQoL-6D instrument contains 20 items across the six dimensions of independent living, social and family relationship, mental health, coping, pain and sense. Health utility scores were calculated using a published algorithm for adolescents (available at https://www.aqol.com.au/index.php/aqolinstruments?id=92), using Australian population preference weights.

2.2.2 Candidate predictors

Data from six measures of psychological distress (one measure), depressive (two measures) and anxiety (three measures) symptoms were used as candidate predictors to construct TTU models. These measures were selected as they are widely used in clinical mental health services or clinically relevant to the profiles of young people seeking mental health care.

The Kessler Psychological Distress Scale (K6; [10]) was used to measure psychological distress over the last 30 days. It includes six items (nervousness, hopelessness, restlessness, sadness, effort, and worthlessness) of the 10 item version of this measure, K10. Individual items use a five-point frequency scale that spans from 0 ("none of the time") to 4 ("all of the time").

The Patient Health Questionnaire-9 (PHQ-9; [11]) and Behavioural Activation for Depression Scale (BADS; [12]) were used to measure degree of depressive symptomatology. PHQ-9 includes nine questions measuring the frequency of depressive thoughts (including self-harm/suicidal thoughts) as well as associated somatic symptoms (e.g., sleep disturbance, fatigue, anhedonia, appetite, psychomotor changes) in the past two weeks. PHQ-9 uses a four-point frequency scale ranging from 0 ("Not at all") to 3 ("Nearly every day"). For the PHQ-9 a total score is derived (0-27) with higher scores depicting greater symptom severity. BADS measures a range of behaviours (activation, avoidance/rumination, work/school impairment as well as social impairment) reflecting severity of depression. BADS includes 25 questions on depression activated behaviours over the past week, scored on a seven-point scale ranging from 0 ("Not at all") to 6 ("Completely"). A total score is derived for the BADS (0-150) as well as subscale scores, with higher scores indicating greater activation (or impairment on the social impairment subscale).

The Generalised Anxiety Disorder Scale (GAD-7; [13]), Screen for Child Anxiety Related Disorders (SCARED; [14]) and Overall Anxiety Severity and Impairment Scale (OASIS; [15]) were used to measure anxiety symptoms. GAD-7 measures symptoms such as nervousness, worrying and restlessness, over the past two weeks using seven questions, with a four-point frequency scale ranging from 0 ("Not at all") to 3 (Nearly every day"). A total score is calculated with scores ranging from 0 to 21 and higher scores indicating more severe symptomatology. SCARED is an anxiety screening tool designed for children and adolescents which can be mapped directly on specific Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) anxiety disorders including generalised anxiety disorder, panic disorder, separation anxiety disorder and social phobia. It includes 41 questions on a three-point scale of 0 ("Not true or hardly ever true"), 1 "Somewhat True or Sometimes True" and 2 ("Very true or often true") to measure symptoms over the last three months. A total score is derived with scores ranging from 0-82, with higher scores indicative of the presence of an anxiety disorder. The OASIS was developed as a brief questionnaire to measure severity of anxiety and impairment in clinical populations. The OASIS includes five questions about frequency and intensity of anxiety as well as related impairments such as avoidance, restricted activities and problems with social functioning over the past week. Total scores range from 0-20 with higher scores depicting more severe symptomatology.

2.2.3 Population characteristics

We collected self-reported measures of demographics (age, gender, education and employment status, languages spoken at home and country of birth). We also collected clinician or research interviewer assessed measures of mental health including primary diagnosis, clinical stage [16] and functioning (measured by the Social and Occupational Functioning Assessment Scale (SOFAS) [17]).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.3 Procedures

The study was approved by the University of Melbourne's Human Research Ethics Committee, and Human Ethics and Advisory Group (1443342.1 and 1645367.1). Eligible participants were recruited by trained research assistants and written consent was obtained from the young person and a parent/guardian if the participant was aged <18 years.

Participants responded to the questionnaire via a tablet device and participants' clinical characteristics were obtained from clinical records and research interview. At three-months post-baseline, participants were contacted in person or by telephone, to complete a 3-month follow-up assessment.

2.4 Statistical analysis

Basic descriptive statistics were used to characterise the cohort in terms of baseline demographics and clinical variables. Pearson's Product Moment Correlations (r) were used to determine the relationships between candidate predictors and the AQoL-6D utility score.

2.4.1 TTU regression models

As AQoL-6D utility score is normally left skewed and constrained between 0 and 1, ordinary least squares (OLS) models with different types of outcome transformations (such as log and logit) have been previously used in TTU regression [3]. Similarly, generalised linear models (GLMs) address this issue via modelling the distribution of the outcome variable and applying a link function between the outcome and linear combination of predictors [18].

We compared predictive performance of a range of models predicting AQoL-6D utility scores using the candidate predictor that had the highest Pearson correlation coefficient with utility scores among the six candidate predictors. The models compared include OLS regression with log, logit, log-log (f(y) = -log(-log(y))) and clog-log (f(y) = -log(1-y))) transformation; GLM using Gaussian distribution with log link; and GLM using Beta distribution with logit and clog-log link. Ten-fold cross-validation was used to compare model fitting using training datasets and predictive ability using testing datasets using three indicators including R^2, root mean square error (RMSE) and mean absolute error (MAE) [19,20].

To evaluate whether our candidate predictors were able to independently predict utility scores, we established multivariate prediction models using baseline data with the candidate predictor and a range of other risk factors including participants' age, gender, clinical stage, cultural and linguistic diversity, education and employment status, primary diagnosis, region of residence (whether metropolitan - based on location of attending service) and sexual orientation. Functioning (as measured by SOFAS), considered an important clinical outcome, was also included in the model to evaluate whether it can jointly predict utility with clinical symptom measurements.

2.4.2 Candidate predictor comparison

Two steps were used to look at the usefulness of the candidate predictors. First, we used a random forest model including all six candidate predictors. This method was used as anxiety and depression measurements are highly collinear, making it difficult to compare these candidate predictors jointly using one regression model. Random forest models commonly used in feature selection provide flexible methods for comparing correlated predictors' relative 'importance' (loss of accuracy from random permutation of the predictor) for the overall prediction model [21]. Second the predictive performance of candidate predictors using selected TTU regression model were compared using 10-fold cross-validation. This procedure helped us to directly evaluate the independent predictive ability of different candidate predictors.

2.4.3 Methods to evaluate the ability of measures to predict longitudinal change in health utility

After identifying the best TTU regression model(s), we established longitudinal models to evaluate the ability to predict change. This was achieved using generalised linear mixed- effect models (GLMM) including both

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

the baseline and follow-up data. The detailed model is specified in the following equation:

$$g(U_{i,j}) = (\beta_0 + b_i) + \beta_{baseline} \times S_{i,baseline} + \beta_{change} \times \Delta S_{i,j} + \epsilon_{i,j} \tag{1}$$

$g()$ is the link function of the model; $U_{i,j}$ is AQoL-6D utility score of individual i in observation j; $S_{i,baseline}$ is the baseline distress/depression/anxiety score for individual i and $\Delta S_{i,j}$ is the score change from the baseline for individual i at observation j. We used β_0 to represent fixed intercept, b_i to represent the random intercept for individual i (controlling for clustering at individual level) and $\epsilon_{i,j}$ to represent the random error. Hence for baseline observations $\Delta S_{i,j} = 0$; and at follow-up $\Delta S_{i,j} = S_{i,follow-up} - S_{i,baseline}$. With this parameterisation, $\beta_{baseline}$ can be interpreted as between person association and β_{change} as within person association. When $\beta_{baseline} = \beta_{change}$, Equation 1 can be generalised to:

$$g(U)_{i,j} = (\beta_0 + b_i) + \beta \times S_{i,j} + \epsilon_{i,j} \tag{2}$$

for both baseline and follow-up observations. The discrepancy between $\beta_{baseline}$ and β_{change} can be interpreted as bias of estimating longitudinal predictive score changes within individual using cross-sectional score difference between individuals.

Bayesian linear mixed models were used to avoid common convergence problems in frequentist tools [22]. Linear mixed effect model (LMM) can be fitted in the same framework with Gaussian distribution and identify link function. Clustering at individual level is controlled via including random intercepts. Model fitting was evaluated using Bayesian R^2 [23].

2.4.4 Secondary analyses

We repeated the previous steps to develop additional TTUs - a set of models that used SOFAS as an independent predictor (Secondary Analysis A) and a set of models that combined anxiety and depression predictors (Secondary Analysis B).

2.4.5 Toolkit development

We undertook all our analyses using ***R*** 4.0.2 [24]. We used a wide range of third- party code libraries in the analysis and reporting (see Supplementary Information, Table A.5). We then rewrote our analysis and reporting algorithms as R packages so that they can be used by others as tools for predicting QALYs, replicating this study and developing TTUs with different utility measures and predictors. Where it is not feasible to publicly release study data synthetic replication datasets can be useful [25]. We created such a dataset to facilitate others trialing our toolkits and rerunning our study algorithm.

3 Results

3.1 Cohort characteristics

Participants characteristics at baseline and follow-up are displayed in Table 1. This study included 1068 out of the 1107 participants with complete AQol-6D data. This cohort predominantly comprised individuals with anxiety/depression (76.7%) at early (prior to first episode of a serious mental disorder) clinical stages (91.7%). Participant ages ranged between 12-25 with a mean age of 18.13 (SD = 3.26).

There were 643 participants (60.2%) who completed AQol-6D questions at the follow-up survey three months after baseline assessment.

3.2 AQol-6D and candidate predictors

Distribution of AQol-6D total utility score and sub-domain scores are displayed in Figure 1, the mean utility score at baseline is 0.59 (SD = 0.24) and 0.68 (SD = 0.24) at follow-up. Distribution of candidate predictors,

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1: Participant characteristics

		Baseline		Follow-Up	
		(N =	1068)	(N =	643)
Age	Mean (SD)	18.13	(3.26)	18.19	(3.25)
	Median (Q1 Q3)	18.00	(16.00 20.00)	18.00	(16.00 21.00)
	Min - Max	12.00	25.00	12.00	25.00
	Missing	0.00		0.00	
Relationship Status	Not in relationship	695.00	(66.70%)	426.00	(68.27%)
	In relationship	347.00	(33.30%)	198.00	(31.73%)
	Missing	26.00		19.00	
Education and Employment Status	Studying only	405.00	(39.09%)	247.00	(39.71%)
	Working only	167.00	(16.12%)	91.00	(14.63%)
	Studying and working	305.00	(29.44%)	193.00	(31.03%)
	Not studying or working	159.00	(15.35%)	91.00	(14.63%)
	Missing	32.00		21.00	
Primary Diagnosis	Depression	182.00	(17.93%)	108.00	(17.31%)
	Anxiety	264.00	(26.01%)	181.00	(29.01%)
	Depression and Anxiety	332.00	(32.71%)	188.00	(30.13%)
	Other	237.00	(23.35%)	147.00	(23.56%)
	Missing	53.00		19.00	
Clinical Stage	0-1a	625.00	(60.27%)	456.00	(72.04%)
	1b	326.00	(31.44%)	131.00	(20.70%)
	2-4	86.00	(8.29%)	46.00	(7.27%)
	Missing	31.00		10.00	

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 2: Candidate predictors distribution parameters and correlations with AQoL-6D utility

		Baseline		Follow-Up		
		(N =	1068)	(N =	643)	p
Behavioural Activation for Depression Scale (0-150)	Mean (SD)	78.16	(24.82)	89.36	(24.43)	0.00
	Missing	10.00		2.00		0.00
	Correlation with AQOL-6D	0.66		0.66		0.00, 0.00
Generalised Anxiety Disorder Scale (0-21)	Mean (SD)	10.38	(5.66)	7.95	(5.46)	0.00
	Missing	6.00		2.00		0.00
	Correlation with AQOL-6D	-0.65		-0.69		0.00, 0.00
Kessler Psychological Distress Scale (6 Dimension) (0-24)	Mean (SD)	12.16	(5.76)	9.81	(5.87)	0.00
	Missing	4.00		2.00		0.00
	Correlation with AQOL-6D	-0.63		-0.63		0.00, 0.00
Overall Anxiety Severity and Impairment Scale (0-20)	Mean (SD)	8.06	(4.72)	6.29	(4.34)	0.00
	Missing	7.00		1.00		0.00
	Correlation with AQOL-6D	-0.69		-0.71		0.00, 0.00
Patient Health Questionnaire (0-27)	Mean (SD)	12.84	(6.62)	9.84	(6.48)	0.00
	Missing	4.00		5.00		0.00
	Correlation with AQOL-6D	-0.74		-0.78		0.00, 0.00
Screen for Child Anxiety Related Disorders (0-82)	Mean (SD)	34.24	(17.85)	28.83	(17.83)	0.00
	Missing	7.00		2.00		0.00
	Correlation with AQOL-6D	-0.63		-0.63		0.00, 0.00

BADS, GAD-7, K6, OASIS, PHQ-9 and SCARED, are summarised in Table 2. PHQ-9 was found to have the highest correlation with utility score both at baseline and follow-up followed by OASIS and BADS; baseline and follow-up SCARED was found to have the lowest correlation coefficients with utility score although all correlation coefficients can be characterised as being strong.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

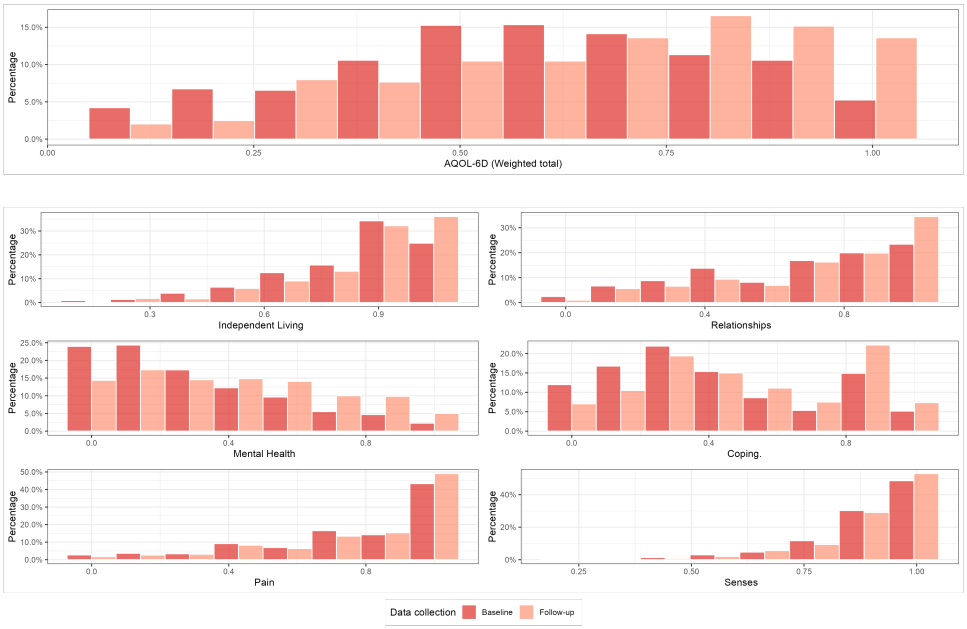

Figure 1: Distribution of AQoL-6D domains

3.3 TTU regression model performance

The 10-fold cross-validated model fitting index from TTU models using PHQ9 are reported in Table A.1 in the Supplementary Material. Both training and testing R^2, RMSE and MAE were comparable between models selected, and GLM using Gaussian distribution and log link had the highest predictive performance. The best OLS model was found to be either no transformation, log transformation or clog-log transformation. Model diagnoses (such as heteroscedasticity, residual normality) suggested better model fit of the clog-log transformed model, as the distribution clog-log transformed utility are closest to normal distribution among all transformation methods. Another benefit of the clog-log model is that the predicted utility score will be constrained with an upper bound of 1, thus preventing out of range prediction. Therefore, both GLM with Gaussian distribution and log link and OLS with clog-log transformation were selected for further evaluation. Predictive ability of each candidate predictor using baseline data were also compared using 10-fold cross-validation.

As shown in Table A.2, PHQ9 had the highest predictive ability followed by OASIS, BADS, GAD7 and K6. SCARED had the least predictive capability. This is consistent with random forest feature selection model, when all of the measurements were used in one model to predict utility score, PHQ9 was found to be the most 'important' predictor (see Figure A.1). The confounding effect of other participant characteristics were also evaluated when using the candidate predictors in predicting utility score (results not shown). Using the baseline data, SOFAS was found to independently predict utility scores in models for all six candidate predictors *($p<0.005$)*. No other confounding factor was identified for the either predictor prediction model; sex at birth was found to be a confounder for K6 model *($p<0.01$)*. A few other confounders, including primary diagnosis, clinical staging and age were identified as weakly associated with utility in TTU models using anxiety and depression measurements other than PHQ-9. Considering many of these factors are unlikely to change over three months, they were not evaluated in the mixed effect models.

3.4 Longitudinal TTU regression model

Regression coefficients of the baseline score and score changes (from baseline to follow-up) estimated in individual GLMM and LMM models are summarised in Table 3. Bayesian R^2 from each model is reported. Modelled residual standard deviations (SDs) were also provided to support simulation studies which need to

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

capture individual level variation. In GLMM and LMM models, the prediction models using OASIS and PHQ-9 respectively had the highest R^2 (0.66 and 0.76) and lowest estimated residual SD. R^2 were above 0.7 for all LMM models and above 0.6 for all GLMM models except for the K6 model. Variance of the random intercept was comparable with the residual variance. A detailed summary of all models from the primary analysis is available in the online results repository (see "Availability of data and materials").

Table 3: Estimated coefficients from longitudinal TTU models for candidate predictors

	GLMM with Gaussian distribution and log link					LMM with clog-log transformation				
Parameter	Estimate	SE	95CI	R2	Sigma	Estimate	SE	95CI	R2	Sigma
PHQ9 model				0.66	0.14				0.76	0.41
SD (Intercept)	0.11	0.01	0.09, 0.14			0.36	0.02	0.32, 0.39		
Intercept	0.01	0.01	-0.01, 0.04			1.12	0.03	1.06, 1.19		
PHQ9 baseline	-4.50	0.11	-4.72, -4.29			-9.62	0.23	-10.08, -9.16		
PHQ9 change	-3.87	0.15	-4.16, -3.58			-8.10	0.30	-8.68, -7.51		
OASIS model				0.68	0.13				0.76	0.41
SD (Intercept)	0.18	0.01	0.16, 0.20			0.43	0.02	0.40, 0.46		
Intercept	-0.09	0.01	-0.11, -0.06			0.91	0.03	0.84, 0.97		
OASIS baseline	-5.91	0.18	-6.26, -5.55			-12.51	0.36	-13.22, -11.80		
OASIS change	-5.61	0.24	-6.09, -5.15			-12.03	0.48	-12.97, -11.11		
BADS model				0.63	0.14				0.73	0.43
SD (Intercept)	0.17	0.01	0.15, 0.19			0.44	0.02	0.41, 0.48		
Intercept	-1.39	0.03	-1.45, -1.33			-1.89	0.06	-2.01, -1.78		
BADS baseline	1.06	0.03	1.00, 1.13			2.29	0.07	2.15, 2.43		
BADS change	0.85	0.04	0.76, 0.93			1.85	0.09	1.68, 2.02		
SCARED model				0.62	0.15				0.71	0.45
SD (Intercept)	0.18	0.01	0.16, 0.20			0.46	0.02	0.42, 0.49		
Intercept	-0.08	0.02	-0.11, -0.04			0.92	0.04	0.84, 1.00		
SCARED baseline	-1.38	0.05	-1.48, -1.28			-2.95	0.10	-3.15, -2.75		
SCARED change	-1.46	0.08	-1.61, -1.31			-3.29	0.16	-3.60, -2.97		
K6 model				0.58	0.15				0.71	0.45
SD (Intercept)	0.16	0.01	0.13, 0.19			0.46	0.02	0.43, 0.50		
Intercept	-0.03	0.02	-0.07, 0.00			1.03	0.04	0.95, 1.12		
K6 baseline	-4.23	0.14	-4.52, -3.95			-9.26	0.32	-9.90, -8.62		
K6 change	-3.55	0.20	-3.93, -3.17			-7.48	0.37	-8.19, -6.76		
GAD7 model				0.62	0.15				0.73	0.43
SD (Intercept)	0.16	0.01	0.13, 0.18			0.44	0.02	0.41, 0.48		
Intercept	-0.08	0.01	-0.10, -0.05			0.92	0.04	0.85, 0.99		
GAD7 baseline	-4.64	0.15	-4.93, -4.35			-9.87	0.32	-10.50, -9.25		
GAD7 change	-4.21	0.19	-4.60, -3.83			-8.81	0.38	-9.57, -8.08		

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The coefficients of score change from baseline were generally estimated to be lower compared with coefficients of baseline score (except for SCARED). This suggests possible overestimation of utility change using the estimates derived from cross-sectional studies. The ratio between two coefficients ($\beta_{change}/\beta_{baseline}$) is 0.9 for K6, 0.82 for depression measurements and 0.8 or over for anxiety measurements.

Distribution of observed and predicted utility scores and their association from GLMM (Gaussian distribution with log link) and LMM (c-loglog transformed)) using OASIS and PHQ-9 respectively are plotted in Figure 2. Compared with GLMM, the predicted utility scores from the LMM model converge better to the observed distribution and provide better estimations at the tail of the distribution. When the observed utility scores were low, the predicted utility were too high in GLMM model, see Figure 2 (B). The observed and predicted distributions of utility scores for other anxiety and depression measurements were similar from LMM models. However, the predicted distributions depart substantially from the target distribution for LMM models, with low coverage in utility scores below 0.3 and also prediction out of range (over 1).

Our primary analysis also evaluated models with SOFAS at baseline and SOFAS change from baseline added to Psychological distress, Depression and Anxiety predictors (see Tables A.3 and A.4). SOFAS scores were generally found to be associated with utility scores when controlling for anxiety and depression symptom measurements in longitudinal models.

The secondary analysis where SOFAS is the sole predictor resulted in models with slightly lower R^2 than all primary analysis models. Adding the PHQ-9 depression measure to each anxiety measure predictor did not notably improve the performance of these models. Results from the secondary analyses are available in the online results repository (see "Availability of data and materials").

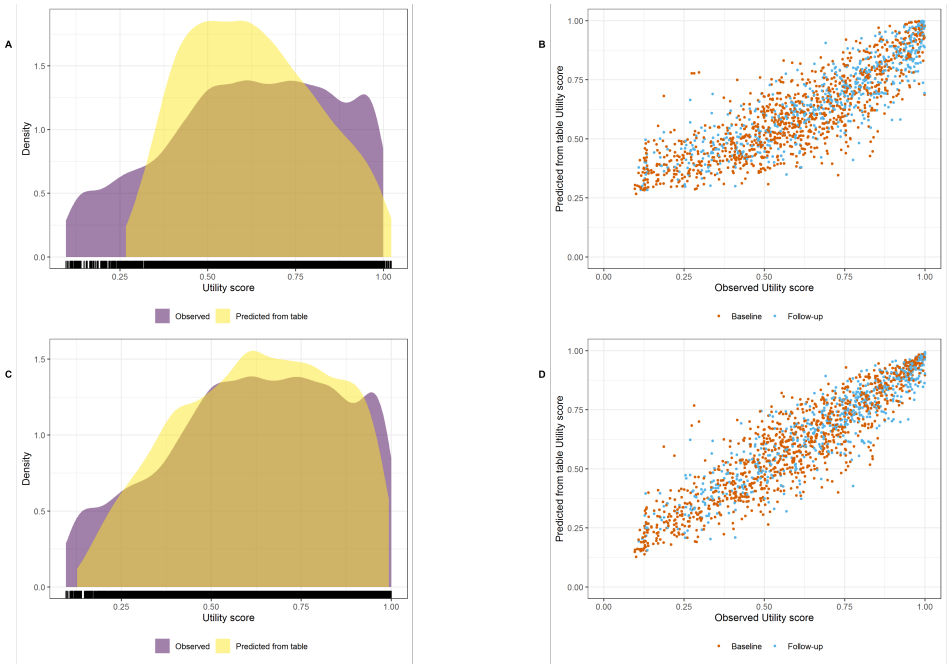

Figure 2: Comparison of observed and predicted AQoL-6D utility score from longitudinal TTU of PHQ-9 (A) Density plots of observed and predicted utility scores (GLMM with Gaussian distribution and log link) (B) Scatter plots of observed and predicted utility scores by timepoint (GLMM with Gaussian distribution and log link) (C) Density plots of observed and predicted utility scores (LMM with clog-log transformation) (D) Scatter plots of observed and predicted utility scores by timepoint (LMM with clog-log transformation))

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

3.5 Toolkits for predicting QALYs and modelling additional TTUs

We created an online results data-repository and three R packages to facilitate easy access to and application of study outputs and replication of study methods. See "Availability of data and materials" for details of where these resources (and supporting documentation) can be accessed.

4 Discussion

MAUIs are largely absent in routine data collection in clinical mental health services. This gap means that it can be difficult for researchers, service planners and service commissioners to derive much economic insight from the often-rich outcome data that is collected in administrative and treatment evaluation datasets. Existing TTU algorithms may not appropriately predict longitudinal change in utility weights especially in help-seeking young people. Our study addresses this important gap and is the first to evaluate longitudinal mapping ability between affective symptom measurements and health utility in a cohort of help seeking young people.

Although there is encouraging evidence about the quality, effectiveness and cost-effectiveness of youth mental health service innovations worldwide [26][27], the public health and economic returns from large scale systemic reforms to support better mental health in young people still needs to be better understood [28]. Our study contributes to this goal by developing tools that can extract additional economic insights into existing mental health datasets by facilitating prediction of QALYs with our TTU algorithms and supporting the development of additional TTU algorithms by other researchers.

By helping to translate measures commonly collected in youth mental health services to QALYs, our TTU algorithms enable greater use of cost-utility analyses (CUAs) - which unlike alternative economic evaluation types (e.g., Cost Consequence Analysis and Cost-Effectiveness Analysis using measures other than health utility) has commonly understood willingness to pay benchmarks and facilitates comparison of the value for money claims of interventions from different illness groups. In practical terms, CUAs can help a decision-maker assess the competing economic claims of an intervention for depression compared to an intervention in anxiety or determine whether a budget might be more efficiently allocated by disinvesting from some established interventions in physical health to fund expanded access to specified mental health services.

The vulnerability of young people to poorer mental health arising from major social and economic disruptions such as the COVID-19 pandemic is an urgent reminder that we need to better use existing datasets to provide real-time decision support. As many youth mental health services routinely collect data on at least one of our six candidate predictors and the measure of functioning (SOFAS) included in our models, the TTU algorithms we developed in this study may have widespread applicability. Importantly, our TTUs were developed in a clinical sample of 12-25 year olds, using the adolescent version of the AQoL-6D. We were able to independently predict adolescent AQoL-6D from each of the six candidate measures we assessed, with PHQ-9 having the best predictive performance. Predictive performance was improved when adding SOFAS as an additional predictor or confound to each model; SOFAS also performed well as an independent predictor. These results may be useful for service system planners in helping to prioritise which measures should be included in routine data collection. Although direct measurement of health utility with measures such as the ReQoL [29] may be feasible in some mental health services, including clinical measures that can also map to health utility may be an attractive alternative.

A key feature of QALYs is their longitudinal dimension - health utilities are weighted and aggregated based on the time spent in varying health states. Our results suggest that psychological distress, depression and anxiety measurements explain the variations of health utility and cross-sectional variations can be used to approximate the longitudinal change in this cohort. However, a finding of our study is that TTU algorithms developed from cross-sectional data may over-estimate these changes, introducing bias into QALY predictions (overestimating QALYs for populations whose health utility improves over time, underestimating QALYS for those with deteriorating mental health). We have therefore identified scaling factors that can be used to adjust predictions from between-person to within-person predictions.

Key strengths of our study include the novelty of our clinical youth mental health study sample, the use of

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

clinically relevant and frequently collected outcome measures as predictors, the appropriateness and range of statistical methods deployed, the comparison of within-person and between-person differences in health utility weight predictions and highly replicable, publicly disseminated study algorithms. We acknowledge limitations that our data pertained to a single country, and we explored only one MAUI-derived utility weight. We did not examine some potential predictors that may be more common in some mental health services (for example we explored K6, as opposed to the expanded, and commonly used measure, the K10).

However, using utility weight input data derived from the same country as that to which an analysis pertains may be relatively unimportant [30], particularly when the MAUI is well suited to the relevant health condition (as is the case with AQoL and mental health [9]). Furthermore, our R packages should help make it relatively straightforward for others to replicate our study algorithm in different samples (non-Australian, non-clinical and/or non-youth populations) and generalise our methods to developing TTU algorithms that use different predictors (other clinical, functioning and demographic measures) and other utility measures (e.g., EQ-5D). Clinical trial datasets, which now usually collect MAUIs, could provide rich opportunities for applying our algorithm to develop and test new TTU algorithms.

We have distributed a large body of study outputs as freely available open science resources - principally freely accessible TTU datasets and open source software. By doing so we hope to make it easier to access and appropriately and consistently use study results. Open science resources also provide a valuable opportunity for other researchers to contribute refinements and extensions so that the usefulness of our study algorithm improves with time.

5 Conclusions

We have found that it is possible to predict both within-person and between-person differences in adolescent AQOL-6D utility weights from measures routinely collected in youth mental health services. Using TTU algorithms developed from cross-sectional data to predict longitudinal changes in health utility may slightly over-estimate these changes. The TTU algorithms we have developed, and the scaling factors we identified to adjust predictions from between-person TTU algorithms to within-person predictions, can help inform resource allocation decisions relating to the mental health of young people. Our toolkits also provide a basis for future research that extends our work with additional TTU algorithms.

Availability of data and materials

Detailed results in the form of catalogues of the TTU models produced by this study and other supporting information are available in the results repository https://doi.org/10.7910/DVN/DKDIB0. Tools for finding and using the TTU models appropriate for use with new prediction datasets are available as part of the youthu R package (https://ready4-dev.github.io/youthu). The youthvars R package (https://ready4-dev.github.io/youthvars/) provides a number of tools helpful for replicating this study (including a synthetic dataset) while TTU (https://ready4-dev.github.io/TTU/) has tools for both replicating the study and generalising our algorithms to develop TTU algorithms with other utility measures and predictors.

Ethics approval

The study was reviewed and granted approval by the University of Melbourne's Human Research Ethics Committee, and the local Human Ethics and Advisory Group (1645367.1).

Funding

This study was funded by the National Health and Medical Research Council (NHMRC, APP1076940), Orygen and headspace.

Conflict of Interest

None declared.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. MacKillop E, Sheard S. Quantifying life: Understanding the history of quality-adjusted life-years (qalys). Social Science & Medicine. 2018;211: 359–366. doi:https://doi.org/10.1016/j.socscimed.2018.07.004

2. Neumann PJ, Goldie SJ, Weinstein MC. Preference-based measures in economic evaluation in health care. Annual Review of Public Health. 2000;21: 587–611. doi:10.1146/annurev.publhealth.21.1.587

3. Mortimer D, Segal L. Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into qaly-weights. Medical decision making. 2008;28: 66–89.

4. Henry JD, Crawford JR. The short-form version of the depression anxiety stress scales (dass-21): Construct validity and normative data in a large non-clinical sample. British journal of clinical psychology. Wiley Online Library; 2005;44: 227–239. doi:https://doi.org/10.1348/014466505X29657

5. Mihalopoulos C, Chen G, Iezzi A, Khan MA, Richardson J. Assessing outcomes for cost-utility analysis in depression: Comparison of five multi-attribute utility instruments with two depression-specific outcome measures. The British Journal of Psychiatry. 2014;205: 390–397.

6. Furber G, Segal L, Leach M, Cocks J. Mapping scores from the Strengths and Difficulties Questionnaire (SDQ) to preference-based utility values. Qual Life Res. 2014;23: 403–411.

7. Filia K, Rickwood D, Menssink J, Gao CX, Hetrick S, Parker A, et al. Clinical and functional characteristics of a subsample of young people presenting for primary mental healthcare at headspace services across australia. Soc Psychiatry Psychiatr Epidemiol. 2021; doi:10.1007/s00127-020-02020-6

8. Richardson JR, Peacock SJ, Hawthorne G, Iezzi A, Elsworth G, Day NA. Construction of the descriptive system for the assessment of quality of life aqol-6D utility instrument. Health and quality of life outcomes. 2012;10: 38. Available: https://hqlo.biomedcentral.com/track/pdf/10.1186/1477-7525-10-38

9. Engel L, Chen G, Richardson J, Mihalopoulos C. The impact of depression on health-related quality of life and wellbeing: Identifying important dimensions and assessing their inclusion in multi-attribute utility instruments. Qual Life Res. 2018;27: 2873–2884. doi:10.1007/s11136-018-1936-y

10. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychological Medicine. 2002;32: 959–976. doi:10.1017/s0033291702006074

11. Kroenke K, Spitzer RL, Williams JB. The phq: Validity of a brief depression severity measure. Journal of general internal medicine. 2001;16: 606–613. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/pdf/jgi_01114.pdf

12. Kanter JW, Mulick PS, Busch AM, Berlin KS, Martell CR. The behavioral activation for depression scale (bads): Psychometric properties and factor structure. Journal of Psychopathology and Behavioral Assessment. 2006;29: 191–202. doi:10.1007/s10862-006-9038-5

13. Spitzer RL, Kroenke K, Williams JB, Lowe B. A brief measure for assessing generalised anxiety disorder: The gad-7. Archives of Internal Medicine. 2006;166: 1092–1097.

14. Birmaher B, Brent DA, Chiappetta L, Bridge J, Monga S, Baugher M. Psychometric properties of the screen for child anxiety related emotional disorders (scared): A replication study. Journal of the American Academy of Child & Adolescent Psychiatry. 1999;38: 1230–1236.

15. Norman SB, Cissell SH, Means-Christensen AJ, Stein MB. Development and validation of an overall anxiety severity and impairment scale (oasis). Depress Anxiety. 2006;23: 245–9. doi:10.1002/da.20182

16. McGorry PD, Hickie IB, Yung AR, Pantelis C, Jackson HJ. Clinical staging of psychiatric disorders: A heuristic framework for choosing earlier, safer and more effective interventions. Aust N Z J Psychiatry. 2006;40: 616–22. doi:10.1111/j.1440-1614.2006.01860.x

17. Goldman HH, Skodol AE, Lave TR. Revising axis v for dsm-iv: A review of measures of social functioning. Am J Psychiatry. 1992;149: 9.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

18. Dobson AJ, Barnett AG. An introduction to generalized linear models. CRC press; 2018.

19. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media; 2009.

20. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai. Montreal, Canada; pp. 1137–1145.

21. Kursa MB, Rudnicki WR. Feature selection with the boruta package. Journal of Statistical Software, Articles. 2010;36: 1–13. doi:10.18637/jss.v036.i11

22. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: A practical guide for ecology and evolution. Trends in ecology & evolution. Elsevier; 2009;24: 127–135.

23. Gelman A, Goodrich B, Gabry J, Vehtari A. R-squared for bayesian regression models. The American Statistician. 2019;73: 307–309. doi:10.1080/00031305.2018.1549100

24. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available: https://www.R-project.org/

25. Nowok B, Raab GM, Dibben C. synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software. 2016;74: 1–26. doi:10.18637/jss.v074.i11

26. Hetrick SE, Bailey AP, Smith KE, Malla A, Mathias S, Singh SP, et al. Integrated (one-stop shop) youth health care: Best available evidence and future directions. Med J Aust. 2017;207: S5–S18. doi:10.5694/mja17.00694

27. Hamilton MP, Hetrick SE, Mihalopoulos C, Baker D, Browne V, Chanen AM, et al. Identifying attributes of care that may improve cost-effectiveness in the youth mental health service system. Med J Aust. 2017;207: S27–S37. doi:10.5694/mja17.00972

28. Alegría M, NeMoyer A, Falgàs Bagué I, Wang Y, Alvarez K. Social determinants of mental health: Where we are and where we need to go. Current Psychiatry Reports. 2018;20: 95–95. doi:10.1007/s11920-018-0969-9

29. 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 in Health. 2021;24: 281–290. doi:https://doi.org/10.1016/j.jval.2020.10.012

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A Appendix

A.1 Additional tables

Table A.1: 10-fold cross-validated model fitting index for different OLS or GLM models for using PHQ9 total scores as predictor with the baseline data

	Training model fit (averaged over 10 folds)			Testing model fit (averaged over 10 folds)		
Model	R2	RMSE	MAE	R2	RMSE	MAE
OLS						
No transformation	0.55	0.16	0.12	0.54	0.16	0.12
Complementary Log Log transformation	0.54	0.16	0.12	0.54	0.16	0.12
Logit transformation	0.51	0.16	0.12	0.51	0.16	0.12
Log transformation	0.50	0.17	0.13	0.49	0.17	0.13
Log Log transformation	0.47	0.17	0.13	0.47	0.17	0.13
GLM						
Gaussian distribution and log link	0.54	0.16	0.13	0.53	0.16	0.13
Beta distribution and complementary log log link	0.55	0.16	0.12	0.55	0.16	0.12
Beta distribution and logit link	0.55	0.16	0.12	0.54	0.16	0.12

* RMSE: Root Mean Squared Error; MAE: Mean Absolute Error

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table A.2: 10-fold cross-validated model fitting index for different candidate predictors estimated using GLM with Gaussian distribution and log link with the baseline data

	Training model fit (averaged over 10 folds)			Testing model fit (averaged over 10 folds)		
Model	R2	RMSE	MAE	R2	RMSE	MAE
PHQ9	0.54	0.16	0.13	0.53	0.16	0.13
OASIS	0.45	0.17	0.14	0.45	0.17	0.14
BADS	0.43	0.18	0.14	0.43	0.18	0.14
GAD7	0.43	0.18	0.14	0.43	0.18	0.14
K6	0.40	0.18	0.14	0.40	0.18	0.14
SCARED	0.40	0.18	0.15	0.39	0.18	0.15

* RMSE: Root Mean Squared Error; MAE: Mean Absolute Error

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table A.3: Estimated coefficients from longitudinal TTU models based on candidate predictors and SOFAS score using LLM (with cloglog transformation)

Parameter*	Estimate	SE	95CI	R2	Sigma
PHQ9 SOFAS model				0.77	0.41
SD (Intercept)	0.35	0.02	0.31, 0.38		
Intercept	0.43	0.13	0.17, 0.68		
PHQ9 baseline	-9.12	0.25	-9.60, -8.61		
PHQ9 change	-7.32	0.34	-7.96, -6.65		
SOFAS baseline	0.96	0.18	0.62, 1.31		
SOFAS change	1.15	0.23	0.70, 1.61		
OASIS SOFAS model				0.77	0.40
SD (Intercept)	0.40	0.02	0.37, 0.44		
Intercept	-0.24	0.13	-0.50, 0.02		
OASIS baseline	-11.52	0.37	-12.26, -10.80		
OASIS change	-10.77	0.50	-11.75, -9.79		
SOFAS baseline	1.62	0.18	1.26, 1.98		
SOFAS change	1.69	0.22	1.24, 2.13		
BADS SOFAS model				0.74	0.43
SD (Intercept)	0.44	0.02	0.40, 0.47		
Intercept	-2.55	0.12	-2.79, -2.31		
BADS baseline	2.07	0.08	1.91, 2.23		
BADS change	1.60	0.09	1.43, 1.78		
SOFAS baseline	1.26	0.20	0.86, 1.66		
SOFAS change	1.53	0.25	1.05, 2.01		
SCARED SOFAS model				0.74	0.43
SD (Intercept)	0.42	0.02	0.38, 0.45		
Intercept	-0.62	0.14	-0.89, -0.35		
SCARED baseline	-2.65	0.10	-2.85, -2.46		
SCARED change	-2.77	0.16	-3.09, -2.44		
SOFAS baseline	2.17	0.19	1.80, 2.54		
SOFAS change	2.34	0.23	1.87, 2.79		
K6 SOFAS model				0.73	0.44
SD (Intercept)	0.44	0.02	0.41, 0.48		
Intercept	-0.29	0.15	-0.58, 0.00		
K6 baseline	-8.16	0.33	-8.81, -7.52		
K6 change	-6.36	0.38	-7.10, -5.64		
SOFAS baseline	1.80	0.20	1.40, 2.18		
SOFAS change	1.99	0.24	1.50, 2.46		
GAD7 SOFAS model				0.74	0.42
SD (Intercept)	0.41	0.02	0.37, 0.44		
Intercept	-0.57	0.13	-0.82, -0.32		
GAD7 baseline	-8.90	0.30	-9.50, -8.30		
GAD7 change	-7.61	0.41	-8.42, -6.81		
SOFAS baseline	2.11	0.18	1.77, 2.46		
SOFAS change	1.86	0.24	1.38, 2.34		

* Calculated as original scores divided by 100

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table A.4: Estimated coefficients from longitudinal TTU models based on individual candidate predictors and SOFAS score using GLM (Gaussian distribution with log link)

Parameter*	Estimate	SE	95CI	R2	Sigma
PHQ9 SOFAS model				0.66	0.14
SD (Intercept)	0.11	0.01	0.08, 0.13		
Intercept	-0.29	0.06	-0.40, -0.18		
PHQ9 baseline	-4.27	0.12	-4.49, -4.04		
PHQ9 change	-3.57	0.17	-3.89, -3.25		
SOFAS baseline	0.42	0.08	0.27, 0.57		
SOFAS change	0.39	0.11	0.17, 0.60		
OASIS SOFAS model				0.68	0.13
SD (Intercept)	0.16	0.01	0.14, 0.18		
Intercept	-0.67	0.06	-0.78, -0.54		
OASIS baseline	-5.46	0.18	-5.81, -5.12		
OASIS change	-5.08	0.25	-5.58, -4.58		
SOFAS baseline	0.83	0.09	0.66, 0.99		
SOFAS change	0.66	0.11	0.45, 0.86		
BADS SOFAS model				0.64	0.14
SD (Intercept)	0.17	0.01	0.15, 0.19		
Intercept	-1.68	0.06	-1.80, -1.57		
BADS baseline	0.96	0.04	0.89, 1.04		
BADS change	0.74	0.05	0.65, 0.83		
SOFAS baseline	0.56	0.09	0.38, 0.74		
SOFAS change	0.59	0.12	0.35, 0.83		
SCARED SOFAS model				0.63	0.14
SD (Intercept)	0.16	0.01	0.14, 0.18		
Intercept	-0.78	0.07	-0.92, -0.66		
SCARED baseline	-1.24	0.05	-1.33, -1.14		
SCARED change	-1.21	0.08	-1.37, -1.05		
SOFAS baseline	1.00	0.09	0.82, 1.17		
SOFAS change	0.96	0.12	0.73, 1.18		
K6 SOFAS model				0.59	0.15
SD (Intercept)	0.15	0.02	0.11, 0.18		
Intercept	-0.64	0.07	-0.78, -0.51		
K6 baseline	-3.74	0.15	-4.04, -3.45		
K6 change	-3.04	0.20	-3.44, -2.64		
SOFAS baseline	0.83	0.09	0.65, 1.01		
SOFAS change	0.77	0.12	0.52, 1.01		
GAD7 SOFAS model				0.62	0.15
SD (Intercept)	0.14	0.01	0.11, 0.16		
Intercept	-0.75	0.06	-0.87, -0.63		
GAD7 baseline	-4.21	0.15	-4.49, -3.92		
GAD7 change	-3.67	0.21	-4.08, -3.25		
SOFAS baseline	0.96	0.08	0.79, 1.12		
SOFAS change	0.74	0.12	0.51, 0.96		

* Calculated as original scores divided by 100

Table A.5: R Packages used in data analysis and reporting

Package	Version	Citation
arsenal	3.6.3	Ethan Heinzen, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson and Gregory Dougherty (2021). arsenal: An Arsenal of 'R' Functions for Large-Scale Statistical Summaries. R package version 3.6.3. https://CRAN.R-project.org/package=arsenal
assertthat	0.2.1	Hadley Wickham (2019). assertthat: Easy Pre and Post Assertions. R package version 0.2.1. https://CRAN.R-project.org/package=assertthat
BCEA	2.3-1.1	Baio et al (2017). Bayesian Cost Effectiveness Analysis with the R package BCEA. Springer, New York, NY. doi: 10.1007/978-3-319-55718-2, URL: http://www.springer.com/us/book/9783319557168/.
betareg	3.1-4	Cribari-Neto F, Zeileis A (2010). "Beta Regression in R." _Journal ofStatistical Software_, *34*(2), 1-24. doi: 10.18637/jss.v034.i02(URL: https://doi.org/10.18637/jss.v034.i02).
boot	1.3-28	Angelo Canty and Brian Ripley (2021). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-28.
Boruta	7.0.0	Miron B. Kursa, Witold R. Rudnicki (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1-13. URL http://www.jstatsoft.org/v36/i11/.
brms	2.15.0	Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01
caret	6.0-88	Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-88. https://CRAN.R-project.org/package=caret
cmdstanr	0.4.0.9000	Jonah Gabry and Rok Cesnovar (2021). cmdstanr: R Interface to 'CmdStan'. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.
cowplot	1.1.1	Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot
dataverse	0.3.8.9000	Will Beasley, Shiro Kuriwaki, Thomas J. Leeper et al. (). dataverse: R Client for Dataverse 4+ Repositories. R package version 0.3.8.9000.
dplyr	1.0.7	Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

enrichwith	0.3.1	https://github.com/ikosmidis/enrichwith>.
eq5d	0.9.0	Fraser Morton and Jagtar Singh Nijjar (2021). eq5d: Methods for Analysing 'EQ-5D' Data and Calculating 'EQ-5D' Index Scores. R package version 0.9.0. https://CRAN.R-project.org/package=eq5d
faux	1.0.0	Lisa DeBruine, (2021). faux: Simulation for Factorial Designs R package version 1.0.0. Zenodo. http://doi.org/10.5281/zenodo.2669586
ggalt	0.4.0	Bob Rudis, Ben Bolker and Jan Schulz (2017). ggalt: Extra Coordinate Systems, 'Geoms', Statistical Transformations, Scales and Fonts for 'ggplot2'. R package version 0.4.0. https://CRAN.R-project.org/package=ggalt
ggfortify	0.4.11	Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.
ggplot2	3.3.5	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
ggpubr	0.4.0	Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
gridExtra	2.3	Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
here	1.0.1	Kirill Müller (2020). here: A Simpler Way to Find Your Files. R package version 1.0.1. https://CRAN.R-project.org/package=here
Hmisc	4.5-0	Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2021). Hmisc: Harrell Miscellaneous. R package version 4.5-0. https://CRAN.R-project.org/package=Hmisc
hutils	1.6.0	Hugh Parsonage (2020). hutils: Miscellaneous R Functions and Aliases. R package version 1.6.0. https://CRAN.R-project.org/package=hutils
knitr	1.33	Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33.
knitrBootstrap	1.0.2	Jim Hester (2018). knitrBootstrap: 'knitr' Bootstrap Framework. R package version 1.0.2. https://CRAN.R-project.org/package=knitrBootstrap
lifecycle	1.0.0	Lionel Henry and Hadley Wickham (2021). lifecycle: Manage the Life Cycle of your Package Functions. R package version 1.0.0. https://CRAN.R-project.org/package=lifecycle

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

lubridate	1.7.10	Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.
magrittr	2.0.1	Stefan Milton Bache and Hadley Wickham (2020). magrittr: A Forward-Pipe Operator for R. R package version 2.0.1. https://CRAN.R-project.org/package=magrittr
MASS	7.3-54	Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
MatchIt	4.1.0	Daniel E. Ho, Kosuke Imai, Gary King, Elizabeth A. Stuart (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, Vol. 42, No. 8, pp. 1-28. URL https://www.jstatsoft.org/v42/i08/
Matrix	1.3-3	Douglas Bates and Martin Maechler (2021). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.3-3. https://CRAN.R-project.org/package=Matrix
matrixcalc	1.0-4	Frederick Novomestky (2021). matrixcalc: Collection of Functions for Matrix Calculations. R package version 1.0-4. https://CRAN.R-project.org/package=matrixcalc
methods	4.0.2	R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
mice	3.13.0	Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.
pacman	0.5.1	Rinker, T. W. & Kurkiewicz, D. (2017). pacman: Package Management for R. version 0.5.0. Buffalo, New York. http://github.com/trinker/pacman
psych	2.1.6	Revelle, W. (2021) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.6,.
purrr	0.3.4	Lionel Henry and Hadley Wickham (2020). purrr: Functional Programming Tools. R package version 0.3.4. https://CRAN.R-project.org/package=purrr
randomForest	4.6-14	A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

readr	1.4.0	Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr
ready4class	0.0.0.9199	Matthew Hamilton and Glen Wiesner (2021). ready4class: Standardised Developer Tools for Creating and Extending Classes for Use as Part of the Ready4 Suite. https://ready4-dev.github.io/ready4class/, https://github.com/ready4-dev/ready4class, https://www.ready4-dev.com/.
ready4fun	0.0.0.9298	Matthew Hamilton and Glen Wiesner (2021). ready4fun: Standardised Function Authoring and Documentation Tools for Use with the Ready4 Suite. https://ready4-dev.github.io/ready4fun/, https://github.com/ready4-dev/ready4fun, https://www.ready4-dev.com/.
ready4show	0.0.0.9035	Matthew Hamilton and Glen Wiesner (2021). ready4show: Standardised Developer Tools for Sharing Insights from Projects Developed with the Ready4 Suite. https://ready4-dev.github.io/ready4show/, https://github.com/ready4-dev/ready4show, https://www.ready4-dev.com/.
ready4use	0.0.0.9133	Matthew Hamilton and Glen Wiesner (2021). ready4use: Standardised Developer Tools for Retrieving and Managing Data in Projects Developed with the Ready4 Suite. https://ready4-dev.github.io/ready4use/, https://github.com/ready4-dev/ready4use, https://ready4-dev.github.io/ready4/.
rlang	0.4.10	Lionel Henry and Hadley Wickham (2020). rlang: Functions for Base Types and Core R and 'Tidyverse' Features. R package version 0.4.10. https://CRAN.R-project.org/package=rlang
rmarkdown	2.9	JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2021). rmarkdown: Dynamic Documents for R. R package version 2.9. URL https://rmarkdown.rstudio.com.
scales	1.1.1	Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

simstudy	0.2.1	Keith Goldfeld and Jacob Wujciak-Jens (2020). simstudy: Simulation of Study Data. R package version 0.2.1. https://CRAN.R-project.org/package=simstudy
stats	4.0.2	R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
stringi	1.6.2	https://stringi.gagolewski.com/>.
stringr	1.4.0	Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr
Surrogate	1.9	Wim Van der Elst, Paul Meyvisch, Alvaro Florez Poveda, Ariel Alonso, Hannah M. Ensor and Christopher J. Weir & Geert Molenberghs (2021). Surrogate: Evaluation of Surrogate Endpoints in Clinical Trials. R package version 1.9. https://CRAN.R-project.org/package=Surrogate
synthpop	1.6-0	Beata Nowok, Gillian M. Raab, Chris Dibben (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), 1-26. doi:10.18637/jss.v074.i11
testthat	3.0.2	Hadley Wickham. testthat: Get Started with Testing. The R Journal, vol. 3, no. 1, pp. 5–10, 2011
tibble	3.1.2	Kirill Müller and Hadley Wickham (2021). tibble: Simple Data Frames. R package version 3.1.2. https://CRAN.R-project.org/package=tibble
tidyr	1.1.3	Hadley Wickham (2021). tidyr: Tidy Messy Data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr
tidyselect	1.1.1	Lionel Henry and Hadley Wickham (2021). tidyselect: Select from a Set of Strings. R package version 1.1.1. https://CRAN.R-project.org/package=tidyselect
truncnorm	1.0-8	Olaf Mersmann, Heike Trautmann, Detlef Steuer and Björn Bornkamp (2018). truncnorm: Truncated Normal Distribution. R package version 1.0-8. https://CRAN.R-project.org/package=truncnorm
TTU	0.0.0.9280	Caroline Gao and Matthew Hamilton (2021). TTU: Transfer to Utility Mapping Algorithm Toolkit. https://ready4-dev.github.io/TTU/, https://github.com/ready4-dev/TTU, https://ready4-dev.github.io/ready4/.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

utils	4.0.2	R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
viridis	0.6.1	Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, and Cédric Scherer (2021). Rvision - Colorblind-Friendly Color Maps for R. R package version 0.6.1.
youthu	0.0.0.9072	Matthew Hamilton and Caroline Gao (2021). youthu: Youth Outcomes to Health Utility. https://ready4-dev.github.io/youthu/, https://github.com/ready4-dev/youthu, https://www.ready4-dev.com/.
youthvars	0.0.0.9058	Matthew Hamilton and Caroline Gao (2021). youthvars: Youth Mental Health Variables Modelling Toolkit. https://ready4-dev.github.io/youthvars/, https://github.com/ready4-dev/youthvars, https://ready4-dev.github.io/ready4/.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A.2 Additional figures

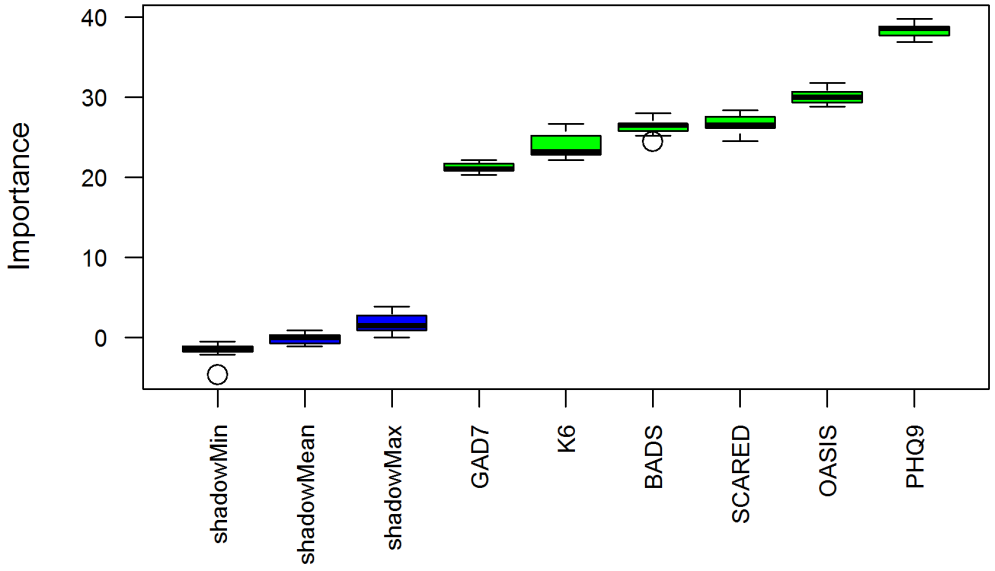

Figure A.1: Variable importance estimated using random forest