medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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.

Mapping psychological distress, depression and anxiety measures to AQoL-6D utility using data from a sample of young people presenting to 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[3] Debra J Rickwood[6,4]
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 identify the best mapping models and predictors for adolescent Assessment of Quality of Life - Six Dimensions (AQOL-6D) utility and assess the ability of mapping models 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 mapping model. Ten-fold cross-validation using R^2, root mean square error (RMSE) and mean absolute error (MAE) were used to compare models and assess predictive ability of six candidate measures of psychological distress, depression and anxiety. 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 performing 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. Mapping models estimated from cross-sectional data can approximate longitudinal change but may slightly bias health utility predictions. *Data:* Replication code, detailed results and guidance on how to apply the models produced by this study are available in the online repository: https://doi.org/10.7910/DVN/DKDIB0.

[†] 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] Faculty of Health, University of Canberra, Australia

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

1 Introduction

To efficiently allocate scarce public resources between competing mental health programs, it is useful to have a common measure of benefit. Quality adjusted life years (QALYs) are generic indices of outcome that inform

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

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 September 13, 2022. 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.

public health policy in many countries [1] and are frequently used in health economic evaluations, including in mental health. The "quality" in QALYs is often measured via the use of multi-attribute utility instruments (MAUIs), where domains of quality of life measured by a questionnaire are weighted using the preferences of people [2]. This approach produces a single health utility weight for each individual for each measured health state, anchored on a scale where 0 represents a state equivalent to 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 used in research studies such as clinical trials and epidemiological surveys, but rarely feature in routine data collection by mental health services. In the absence of direct measurement, mapping analysis has been developed to predict utility weights from standard health status measurements [3,4]. MAUIs such as the Assessment of Quality of Life – 8 dimensions (AQoL-8D [5]) have been shown to be sensitive to psychological distress and depression measures [6] and can be mapped to [7] using measures of psychological distress (measured using Kessler Psychological Distress Scale – 10 items, K10 [8]) and depression and anxiety symptoms (measured using Depression, Anxiety, and Stress Scale – 21 items, DASS-21 [9]). However, it is unclear which mental health measures are the most predictive of health utility and existing algorithms developed for adult [7] or child [10] general populations are of questionable appropriateness for predicting health utility in clinical youth mental health samples.

Currently available mapping algorithms are largely derived from cross-sectional data and assume that the associations between psychological measurements and health utility are time-invariant. Therefore the between-person association (variations in mental health symptoms associated with variations in health utility observed cross-sectionally in a population) can be applied to estimate the within-person association (changes in mental health symptoms associated with changes in health utility over time). However, this time invariant assumption may not be true (for example, health utility measures may be less sensitive to change compared with measures of mental health symptoms).

To inform the development of a simulation model of primary youth mental health services, we aimed to use data from a sample of help-seeking young people to: (i) identify the best mapping regression models to predict adolescent weighted AQoL-6D utility and evaluate the predictive ability of six candidate measures of psychological distress, depression and anxiety; and (ii) assess ability of the mapping algorithms to predict longitudinal (three-month) change.

2 Methods

2.1 Sample and setting

This study forms part of a research program to develop better outcome measures for young people seeking mental health support, and the study sample has previously been described [11]. 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 aged 12-14, more aged 18-20), 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) [11].

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 September 13, 2022. 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; [12]) MAUI. It was selected due to its validity for use in adolescents, the relevance of its domains for a clinical mental health sample [6] and its acceptable participant time-burden. The 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 (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), depression (two measures) and anxiety (three measures) symptoms were used as candidate predictors to construct mapping 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; [8]) 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; [13]) and Behavioural Activation for Depression Scale (BADS; [14]) 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 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.

The Generalised Anxiety Disorder Scale (GAD-7; [15]), Screen for Child Anxiety Related Disorders (SCARED; [16]) and Overall Anxiety Severity and Impairment Scale (OASIS; [17]) 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, sex at birth, 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 [18] and functioning (measured by the Social and Occupational Functioning Assessment Scale (SOFAS) [19]).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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

Eligible participants were recruited by trained research assistants and responded to the questionnaire via a tablet device. 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

We undertook a three-step process to: i) gain insight into our dataset and measures; ii) specify models and assess model performance using cross-sectional data; and iii) specify longitudinal models and assess whether modelling assumptions were met.

2.4.1 Pre-modelling steps

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.

2.4.2 Model specification and assessment (cross-sectional data)

Good practice guidance on mapping studies [4] does not advocate specific model types as the appropriate choice will vary depending on factors that include the health utility measure being mapped to, the applicable health condition and target population, the clinical variables used as predictors and the intended use of the mapping algorithm. Compared to the EQ-5D utility measures most commonly used in the mapping literature [20], AQOL-6D has better dimensional overlap with mental health measures [RN6] and is derived from a greater number of questionnaire items (which can generate more continuous utility score distributions). The types of models we explored reflected these considerations.

We used a cross-section of our dataset (baseline measures only) to explore appropriate type(s) of models to use, compare the relative predictive performance of candidate predictors, and identify other potential risk factors associated with quality of life independent of mental health measures.

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 mapping [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 [21]. Beta regressions, which can be considered a special form of GLMs, are another popular strategy for modelling health utility [22].

We chose to explore OLS, GLM and beta models with commonly adopted transformation algorithms. The models we selected for comparison were OLS regression with log, logit, log-log ($f(y) = -log(-log(y))$) and clog-log ($f(y) = log(-log(1-y))$) transformation; GLM using Gaussian distribution with log link; and beta regression with logit and clog-log link. For each candidate model type, we evaluated the modelling performance and predictive ability of a univariate model using the candidate predictor with the highest Pearson correlation coefficient with utility scores. 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) [23,24]. After identifying the best performing model type, we used 10-fold cross-validation to compare predictive ability of six mental health measures in predicting AQol-6D (one model for each candidate predictor).

To evaluate whether candidate predictors could independently predict utility scores, we added a range of independent variables to each of the six models that included participants' age, sex at birth, clinical stage, cultural and linguistic diversity, education and employment status, primary diagnosis, region of residence

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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.

(whether metropolitan - based on location of attending service) and sexual orientation. Functioning (as measured by SOFAS) was also included in each model to evaluate whether it can jointly predict utility with clinical symptom measurements.

2.4.3 Longitudinal modelling and assumption testing

After identifying the best mapping regression model(s) for predicting between person change, we established longitudinal models to predict within person change. This was achieved using generalised linear mixed- effect models (GLMM) including both the baseline and follow-up data. All records with complete baseline data (with or without follow-up data) were included. 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 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 individuals using cross-sectional score difference between individuals.

Bayesian linear mixed models were used to avoid common convergence problems in frequentist tools [25]. 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 [26].

2.4.4 Secondary analyses

We repeated the previous steps to develop additional mapping algorithms - 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).

2.5 Software

We undertook all our analyses using ***R*** 4.0.5 [27]. We used a wide range of third- party code libraries in the analysis and reporting (see Online Resource, A.5). To aid study replicability and facilitate appropriate use of the mapping models produced by this study we created R programs and code libraries and an online dataset of model metadata and documentation. Where it is not feasible to publicly release study data, synthetic replication datasets can be useful [28]. We created such a dataset and distributed it via an online data repository.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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 Results

3.1 Cohort characteristics

Participants characteristics at baseline and follow-up are displayed in Table 1. This study included 1068 out of the 1073 participants with complete AQoL-6D data. This cohort predominantly comprised individuals with anxiety/depression (76.650%) at early (prior to first episode of a serious mental disorder) clinical stages (91.707%). Participant ages ranged between 12-25 with a mean age of 18.129 (SD = 3.263).

There were 643 participants (60.205%) 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 0.589 (SD = 0.235) and is 0.678 (SD = 0.238) at follow-up. Distribution of candidate predictors, K6, BADS, PHQ-9, GAD-7, OASIS 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.

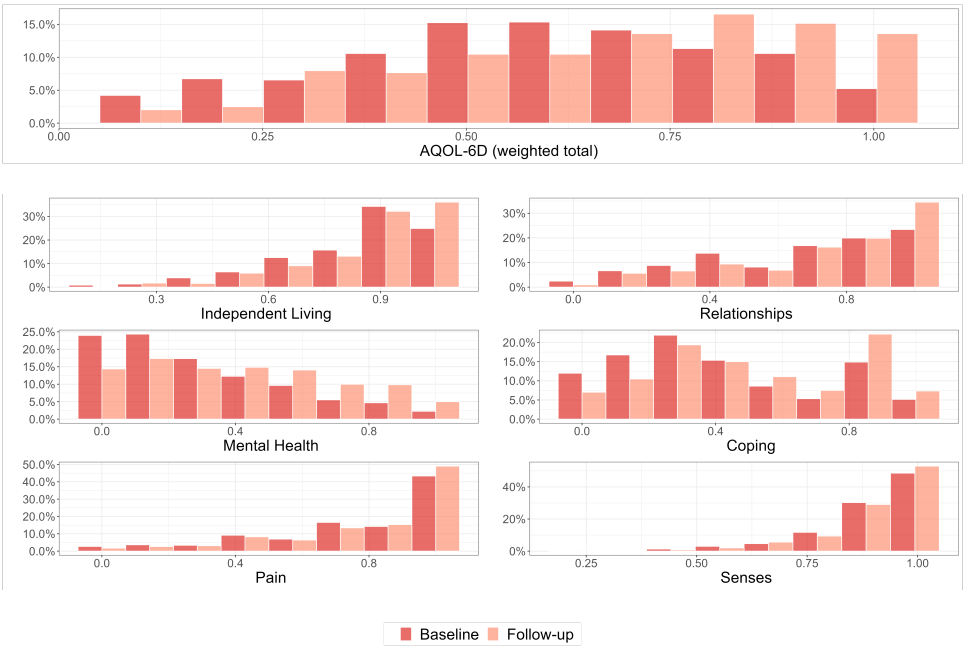

Figure 1: Distribution of AQoL-6D domains

3.3 Regression model performance

The 10-fold cross-validated model fitting index from mapping models using PHQ-9 are reported in Online Resource Table A.1. Both training and testing R^2, RMSE and MAE were comparable between the GLM and beta regression models. OLS model with no transformation or clog-log transformation, GLM with Gaussian distribution and log link and beta regression with logit or cloglog link were found amongst best preforming models. As there were convergence issues with beta regression models, these models were not considered further as they do not show major advantages in predictive performance.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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)	18.129	(3.263)	18.191	(3.249)
	Median (Q1 Q3)	18.000	(16.000 20.000)	18.000	(16.000 21.000)
	Min - Max	12.000	25.000	12.000	25.000
	Missing	0.000		0.000	
Gender	Female	653.000	(62.131%)	398.000	(63.175%)
	Male	359.000	(34.158%)	207.000	(32.857%)
	Other	39.000	(3.711%)	25.000	(3.968%)
	Missing	17.000		13.000	
Relationship status	Not in relationship	695.000	(66.699%)	426.000	(68.269%)
	In relationship	347.000	(33.301%)	198.000	(31.731%)
	Missing	26.000		19.000	
Sexual orientation	Heterosexual	738.000	(71.860%)	431.000	(69.516%)
	Other	289.000	(28.140%)	189.000	(30.484%)
	Missing	41.000		23.000	
Region of residence	Metro	671.000	(62.828%)	382.000	(59.409%)
	Regional	397.000	(37.172%)	261.000	(40.591%)
	Missing	0.000		0.000	
Education and employment status	Studying only	405.000	(39.093%)	247.000	(39.711%)
	Working only	167.000	(16.120%)	91.000	(14.630%)
	Studying and working	305.000	(29.440%)	193.000	(31.029%)
	Not studying or working	159.000	(15.347%)	91.000	(14.630%)
	Missing	32.000		21.000	
Primary diagnosis	Depression	182.000	(17.931%)	108.000	(17.308%)
	Anxiety	264.000	(26.010%)	181.000	(29.006%)
	Depression and Anxiety	332.000	(32.709%)	188.000	(30.128%)
	Other	237.000	(23.350%)	147.000	(23.558%)
	Missing	53.000		19.000	
Clinical stage	0-1a	625.000	(60.270%)	456.000	(72.038%)
	1b	326.000	(31.437%)	131.000	(20.695%)
	2-4	86.000	(8.293%)	46.000	(7.267%)
	Missing	31.000		10.000	
Social and Occupational Functioning Assessment Scale	Mean (SD)	65.182	(9.479)	69.901	(10.082)
	Median (Q1 Q3)	65.000	(60.000 72.000)	70.000	(64.000 78.000)
	Min - Max	29.000	90.000	20.000	91.000
	Missing	30.000		7.000	

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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)	78.158	(24.823)	89.360	(24.435)	0.000
	Missing	10.000		2.000		0.000
	Correlation with AQOL-6D	0.665		0.659		0.000, 0.000
Generalised Anxiety Disorder Scale (0-21)	Mean (SD)	10.382	(5.658)	7.945	(5.460)	0.000
	Missing	6.000		2.000		0.000
	Correlation with AQOL-6D	-0.653		-0.685		0.000, 0.000
Kessler Psychological Distress Scale (6 Dimension) (0-24)	Mean (SD)	12.156	(5.765)	9.811	(5.873)	0.000
	Missing	4.000		2.000		0.000
	Correlation with AQOL-6D	-0.632		-0.634		0.000, 0.000
Overall Anxiety Severity and Impairment Scale (0-20)	Mean (SD)	8.059	(4.718)	6.287	(4.337)	0.000
	Missing	7.000		1.000		0.000
	Correlation with AQOL-6D	-0.688		-0.708		0.000, 0.000
Patient Health Questionnaire (0-27)	Mean (SD)	12.838	(6.616)	9.835	(6.484)	0.000
	Missing	4.000		5.000		0.000
	Correlation with AQOL-6D	-0.742		-0.775		0.000, 0.000
Screen for Child Anxiety Related Disorders (0-82)	Mean (SD)	34.238	(17.852)	28.825	(17.826)	0.000
	Missing	7.000		2.000		0.000
	Correlation with AQOL-6D	-0.635		-0.629		0.000, 0.000

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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.

Model diagnoses (such as heteroscedasticity, residual normality) suggested better model fit of the clog-log transformed OLS model, as the distribution clog-log transformed utility are closest to normal distribution among all transformation methods. Another benefit of the clog-log transformed model is that the predicted utility score will be constrained with an upper bound of 1, thus preventing out of range prediction. Therefore, OLS with clog-log transformation was chosen as the best model for further evaluation As the GLM with Gaussian distribution and log link is also commonly used in mapping studies [20], it is also included to provide easy comparisons with other published work.

PHQ9 had the highest predictive ability followed by OASIS, BADS, GAD7 and K6 (Online Resource Table A.2). SCARED had the least predictive capability. The confounding effect of other participant characteristics were also evaluated when using the candidate predictors in predicting utility score. Using the baseline data, SOFAS was found to independently predict utility scores in models for all six candidate predictors ($p<0.005$). Sex at birth was found to be a confounder for the K6 model ($p<0.01$). A few other confounders, including primary diagnosis, clinical staging and age were identified as weakly associated with utility in mapping 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 mapping models

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 **??**. Bayesian R^2 from each model is reported. Modelled residual standard deviations (SDs) were also provided to support simulation studies which need to capture individual level variation. In GLMM and LMM models, the prediction models using OASIS and PHQ-9 respectively had the highest R^2 (0.681 and 0.762). R^2 was between 0.581 and 0.681 for all GLMMs and between 0.712 and 0.762 for all LMMs.. Variance of the random intercept was comparable with the residual variance.

The coefficients of score change from baseline were generally estimated to be lower compared with coefficients of baseline score (except for SCARED). The mean (across GLMM and LMM models) ratio of the two coefficients ($\beta_{change}/\beta_{baseline}$) is 0.823 for K6, between 0.802 and 0.850 for depression measurements and between 0.900 and 1.086 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 PHQ-9 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 the 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, GLMM models had low coverage in utility scores below 0.3 and also made predictions out of range (over 1).

We also evaluated models with SOFAS at baseline and SOFAS change from baseline added to psychological distress, depression and anxiety predictors (Online Resource, 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.

Detailed summaries of all models from the primary and secondary analyses are available in the online data repository (see “Code and data availability”).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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.

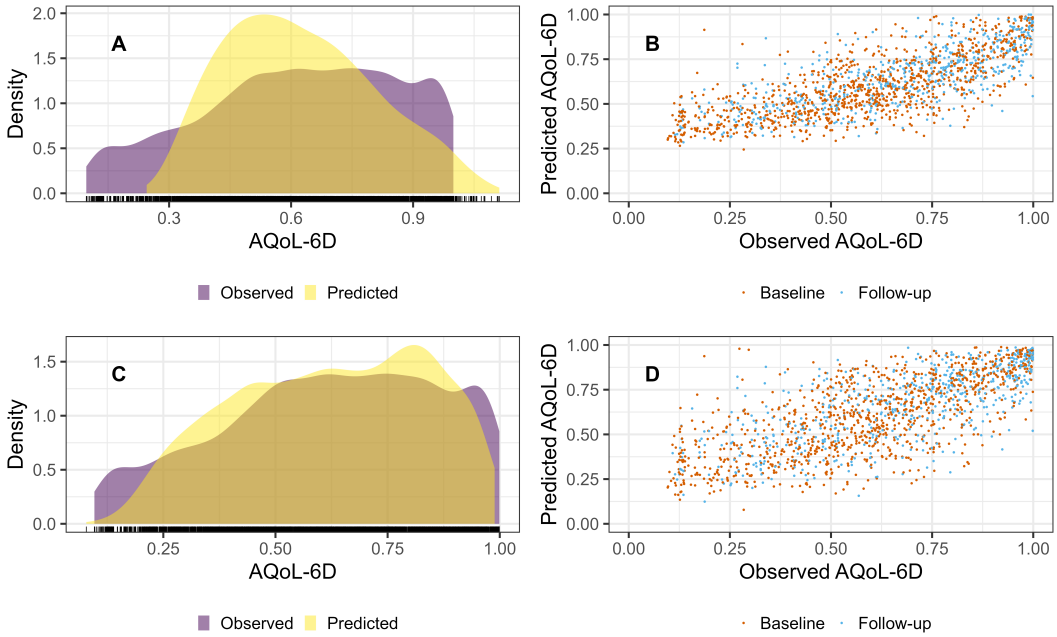

Figure 2: Comparison of observed and predicted AQoL-6D score from longitudinal model using PHQ-9 (A) Density plots of observed and predicted utility scores (GLMM (Gaussian distribution and log link)) (B) Scatter plots of observed and predicted utility scores by timepoint (GLMM (Gaussian distribution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

3.5 Toolkits

To aid study replication, we created a unified R program that performs all steps from raw data ingest, data pre-processing, utility scoring, model specification and testing, creation and documenting of shareable outputs, dissemination of study outputs and authoring of a draft manuscript [29]. To facilitate more generalised use of parts of our study analysis and reporting algorithms, we distributed the code to run program sub-routines as development version R packages for describing and validating input data [30], scoring health utility data [31], specifying and testing mapping models [32] and reporting mapping studies [33].

To facilitate appropriate use of the mapping models we developed we created catalogues with detailed information about the performance of all models generated by our study under multiple types of regime (e.g. whether: mean or sampled parameter values are used; minimum and maximum constraints are applied to prediction results; the original R model objects or just tables of model coefficients are used). We distribute these results in an online data repository [34] along with instructions about our mapping models can be used to make predictions with new data using another R package [35] that we developed.

4 Discussion

Our study was motivated by a desire to incorporate QALY outcomes in a simulation model of primary youth mental health services that we are developing. However, MAUIs are rarely included in routine data collection in clinical mental health services and direct measures of health utility may not exist in many relevant datasets. This gap means that it can be difficult to derive much economic insight from the often-rich outcome data that is collected in administrative and treatment evaluation datasets. Utility mapping is a commonly adopted solution to such problems, but existing mapping algorithms may not appropriately predict longitudinal change

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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.

in utility weights in help-seeking young people. Our study addresses this important gap and is the first to evaluate longitudinal mapping ability between a wide range of 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 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 derivation of QALYs from measures commonly collected in youth mental health services.

Our mapping algorithms enable greater use of cost-utility analyses (CUAs) in studies using youth mental health data collections. Unlike alternative economic evaluation types (e.g., Cost Consequence Analysis and Cost-Effectiveness Analysis using measures other than health utility) CUAs have commonly understood willingness to pay thresholds for outcomes and facilitate 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 it may be efficient to fund expanded access to specified mental health services by redirecting parts of the general health budget.

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 mapping algorithms we developed in this study may have widespread applicability. Importantly, our mapping algorithms were developed in a clinical sample of 12-25 year-olds, using adolescent AQoL-6D weights. 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, relying on 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, for psychological distress and depression measures at least, mapping algorithms developed from cross-sectional data may slightly 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).

Key strengths of our study include the novelty of our clinical youth mental health study sample, the use of 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). We also did not develop age-group specific models which would be a potentially useful extension of this work.

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 and extend our study algorithm in different samples (non-Australian, non-clinical and/or non-youth populations) and generalise our methods to developing mapping 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

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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.

applying our algorithm to develop and test new mapping algorithms.

By distributing study outputs as freely available open science resources we hope to make it easier to access and appropriately and consistently apply study findings. 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. Mapping algorithms developed from cross-sectional data can approximate longitudinal changes in health utility, but may slightly over-estimate these changes. The mapping algorithms we have developed can help inform resource allocation decisions relating to the mental health of young people. Our replication toolkits also provide a basis for future research that extends our work with additional mapping algorithms.

Availability of data and materials

Utility mapping models, instructions on how to apply them and study replication code are distributed as part of the study data repository (https://doi.org/10.7910/DVN/DKDIB0).

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)

Competing interests

None declared.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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. Wailoo AJ, Hernandez-Alava M, Manca A, Mejia A, Ray J, Crawford B, et al. Mapping to estimate health-state utility from non-preference-based outcome measures: An ISPOR good practices for outcomes research task force report. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. United States; 2017;20: 18–27. doi:10.1016/j.jval.2016.11.006
5. Richardson J, Iezzi A, Khan MA, Maxwell A. Validity and reliability of the assessment of quality of life (AQoL)-8D multi-attribute utility instrument. The Patient - Patient-Centered Outcomes Research. 2014;7: 85–96. doi:10.1007/s40271-013-0036-x
6. 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
7. 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.
8. 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
9. 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
10. 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.
11. 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
12. 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
13. 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
14. 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
15. 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.
16. 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.07.21260129; this version posted September 13, 2022. 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.

17. 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

18. 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

19. 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.

20. Mukuria C, Rowen D, Harnan S, Rawdin A, Wong R, Ara R, 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. Applied Health Economics and Health Policy. 2019;17: 295–313. doi:10.1007/s40258-019-00467-6

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

22. Hunger M, Baumert J, Holle R. Analysis of SF-6D index data: Is beta regression appropriate? Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. United States; 2011;14: 759–767. doi:10.1016/j.jval.2010.12.009

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

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

25. 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.

26. 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

27. 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/

28. 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

29. Hamilton M, Gao C. Complete study program to reproduce all steps from data ingest through to results dissemination for a study to map mental health measures to AQoL-6D health utility [Internet]. Zenodo; 2022. doi:10.5281/zenodo.7073428

30. Hamilton M, Gao C. youthvars: Describe and Validate Youth Mental Health Datasets [Internet]. Zenodo; 2022. doi:10.5281/zenodo.7073267

31. Hamilton M, Gao C. Scorz: Score questionnaire item responses [Internet]. Zenodo; 2022. doi:10.5281/zenodo.7073270

32. Hamilton M, Gao C. specific: Specify Candidate Models for Representing Mental Health Systems [Internet]. Zenodo; 2022. doi:10.5281/zenodo.7073273

33. Gao C, Hamilton M. TTU: Implement Transfer to Utility Mapping Algorithms [Internet]. Zenodo; 2022. doi:10.5281/zenodo.7073283

34. Hamilton MP, Gao CX, Filia KM, Menssink JM, Sharmin S, Telford N, et al. Transfer to AQoL-6D Utility Mapping Algorithms [Internet]. Harvard Dataverse; 2021. doi:10.7910/DVN/DKDIB0

35. Hamilton MP, Gao CX. Youthu: Transform youth outcomes to health utility predictions [Internet]. Zenodo; 2022. doi:10.5281/zenodo.7073288