ORIGINAL RESEARCH

How can we estimate QALYs based on PHQ-9 scores? Equipercentile linking analysis of PHQ-9 and EQ-5D

Toshi A Furukawa 1, Stephen Z Levine 2 Claudia Buntrock,3 David D Ebert,4 Simon Gilbody,5 Sally Brabyn,6 David Kessler,6 Cecilia Björkelund,7 Maria Eriksson,7 Annet Kleiboer,4 Annemieke van Straten,4 Heleen Ripper,4 Jesus Montero-Marin,8 Javier Garcia-Campayo,9,10 Rachel Phillips,11 Justine Schneider,12 Pim Cuijpers,4 Eirini Karyotaki4

ABSTRACT

Background Quality-adjusted life years (QALYs) are widely used to measure the impact of various diseases on both the quality and quantity of life and in their economic evaluations. It will be clinically important and informative if we can estimate QALYs based on measurements of depression severity.

Objective To construct a conversion table from the Patient Health Questionnaire-9 (PHQ-9), the most frequently used depression scale in recent years, to the Euro-Qol Five Dimensions Three Levels (EQ-5D-3L), one of the most commonly used instruments to assess QALYs.

Methods We obtained individual participant data of randomised controlled trials of internet cognitive-behavioural therapy which had administered depression severity scales and the EQ-5D-3L at baseline and at end of treatment. Scores from depression scales were all converted into the PHQ-9 according to the validated algorithms. We used equipercentile linking to establish correspondences between the PHQ-9 and the EQ-5D-3L.

Findings Individual-level data from five trials (total N=2457) were available. Subthreshold depression (PHQ-9 scores between 5 and 10) corresponded with values of 0.9–0.8, mild major depression (10–15) with 0.8–0.7, moderate depression (15–20) with 0.7–0.5 and severe depression (20 or higher) with 0.6–0.0. A five-point improvement in PHQ-9 corresponded approximately with an increase in EQ-5D-3L score by 0.03 and a ten-point improvement by approximately 0.25.

Conclusions and Clinical Implications The conversion table between the PHQ-9 and the EQ-5D-3L scores will enable fine-grained assessment of burden of depression at its various levels of severity and of impacts of its various treatments.

INTRODUCTION

Quality-adjusted life years (QALYs) have been increasingly used in general medicine and in psychiatry to evaluate the impact of a disease on both the quantity and quality of life. One QALY is equal to 1 year in perfect health, can range down to zero (death) or may take negative values (worse than death). QALYs can be used to compare the burdens of various diseases, to appreciate the impact of their interventions, to help set priorities in resource allocations across different diseases and interventions and to inform personal decisions.

The representative method to evaluate QALYs is the generic, preference-based measure of health including the Euro-Qol five dimensions (EQ-5D)2 3 and the SF-6D based on Short Form Survey-36 (SF-36).4 5 Of these, the EQ-5D is the most frequently used and is the preferred instrument by the National Institute of Health and Care Excellence in the UK. While the responsiveness of such generic measures to various mental conditions, especially severe mental illnesses, has been questioned, its validity and responsiveness to common mental disorders including depression and anxiety have been generally established.7 8

However, the traditional focus of measurements in mental health has centred mainly on symptoms. Many trials have, therefore, not administered the generic health-related quality of life measures. This has hindered comparison of impacts of mental disorders vis-à-vis other medical conditions on the one hand and also evaluation of values of their interventions on the other.9 10

We have been collecting individual participant-level data from randomised controlled trials of internet cognitive-behavioural therapies (iCBT) for depression,11 several of which administered both symptomatic scales and generic health status scales simultaneously. This study, therefore, attempts to link the depression-specific measure onto the generic measure of health in order to enable estimation of QALYs for depressive states and their changes. Such cross-walking should facilitate assessment of burden of depression at its various severity and of the impacts of its various treatments.

METHODS

Database

We have been accumulating a data set of individual participant data of randomised controlled trials of iCBT among adults with depressive symptoms, as established by specified cut-offs on self-report scales or by diagnostic interviews.11 For this study, we have selected studies that have administered the EQ-5D and depression severity scales at baseline and at end of treatment. We excluded patients if...
they had missing data in either of the two scales at baseline or at endpoint. We excluded studies that focused on patients with general medical disorders (eg, diabetes, glioma) and depressive symptoms.

**Measures**

**EQ-5D-3L**

The EQ-5D-3L comprises five dimensions of mobility, self-care, usual activities, pain/discomfort and anxiety/depression, each rated on three levels corresponding with 1=no problems, 2=some/mild problems or 3=extreme problems/unable to do. This produces $3^5=243$ different health states, ranging from no problem at all in any dimension (11111) to severe problems on all dimensions (33333). Each of these 243 states is provided with a preference-based score, as determined through the time trade-off (TTO) technique in a sample of the general population. In TTO, respondents are asked to give the relative length of time in full health that they would be willing to sacrifice for the poor health states as represented by each of the 243 combinations above. The EQ-5D scores range between 1=full health and 0=death to minus values=worse than death bounded by $-1$. The scoring algorithm for the UK is based on TTO responses of a random sample (n=2997) of noninstitutionalised adults. Over the years, value sets for EQ-5D-3L have been produced for many countries/regions.13–19

**Depression severity scales**

We included any validated depression severity measures. The scale scores were converted into the most frequently used scale, namely, the Patient Health Questionnaire-9 (PHQ-9),20 using the established conversion algorithms14–15 or the Beck Depression Inventory, second edition (BDI-II)15 or the Centre for Epidemiologic Studies Depression Scale (CES-D).16

The PHQ-9 consists of the nine diagnostic criteria items of major depression from the DSM-IV, each rated on a scale between 0 and 3, making the total score range 0–27. The instrument has demonstrated excellent reliability, validity and responsiveness. The cut-offs have been proposed as 0–4, 5–9, 10–14, 15–19 and 20– for no, mild, moderate, moderately severe and severe depression, respectively.12

**Statistical analyses**

We first calculated Spearman correlation coefficients between PHQ-9 and EQ-5D total scores at baseline, at end of treatment and their changes, to establish if the linking is justified. Correlations were considered weak if scores were $<0.3$, moderate if scores were $\geq 0.3$ and $<0.7$ and strong if scores were $\geq 0.7$.17

Correlations $\geq 0.3$ have been recommended to establish linking.18

We then applied the equipercentile linking procedure,19 which identified scores on PHQ-9 and EQ-5D or their changes with the same percentile ranks and allows for a nominal translation from PHQ-9 to EQ-5D by using their percentile values. This approach has been used successfully for scales in depression, schizophrenia or Alzheimer’s disease.14–20–22

We analysed all trials collectively rather than by trial to maximise the sample size, ensure variability in the included populations and attain robust estimates.

We conducted a sensitivity analysis by excluding studies that require the conversion of various depression severity scores into PHQ-9. All the analyses were conducted in R V.4.0.2, with the package equate V.2.0.7.23

**Ethics statement**

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. Ethical approval was not required for this study as it used only deidentified patient data.

**FINDINGS**

**Included studies**

We identified seven RCTs of iCBT (total n=2457), which administered validated depression scales and EQ-5D both at baseline and at endpoint (online supplemental eTable 1). Three studies included only patients with major depressive disorder (MDD), one only patients with subthreshold depression and the remaining three included both. All the studies administered EQ-5D-3L. PHQ-9 scores were converted from the BDI-II in three studies24–26 and from the CES-D in one study.27 The mean age of the participants was 41.8 (SD=12.3) years, 66.0% (1622/2457) were women and they scored 14.0 (5.4) on PHQ-9 and 0.74 (0.20) on EQ-5D at baseline and 9.1 (6.0) and 0.79 (0.21), respectively, at endpoint. When using the standard cut-offs of the PHQ-9,12 2.4% (60/2449) suffered from no depression (PHQ-9 scores <5), 20.2% (492/2449) from subthreshold depression (5≤PHQ-9 scores <10), 33.5% (820/2449) from mild depression (10≤PHQ-9 scores <15), 26.5% (649/2449) from moderate depression (15≤PHQ-9 scores <20) and 17.3% (424/2449) from severe depression (20≤PHQ-9 scores) at baseline.

**Equipercentile linking**

Spearman’s correlation coefficient between the PHQ-9 and the EQ-5D scores was $r=-0.29$ at baseline, increased to $r=-0.50$ after intervention and was $r=-0.38$ for change scores.

**Figure 1** shows the equipercentile linking between PHQ-9 and EQ-5D total scores at baseline and at endpoint. **Figure 2** shows the same between their change scores. Table 1 summarises the correspondences between the two scales.

**Sensitivity analysis**

When we limited the samples to the three studies28–30 that administered PHQ-9 (total n=1375), the linking results were replicated (online supplemental eFigure 1).

---

**Figure 1** PHQ-9 and EQ-5D total scores at baseline and endpoint. EQ-5D, Euro-Qol Five Dimensions; PHQ-9, Patient Health Questionnaire-9.
DISCUSSION
This is the first study to link a depression severity measure with the EQ-5D-3L both for total and change scores. To summarise, subthreshold depression corresponded with EQ-5D-3L index values of 0.9–0.8, mild major depression with 0.8–0.7, moderate depression with 0.7–0.5 and severe depression with 0.6–0.0. A five-point improvement in PHQ-9 corresponded approximately with an increase in EQ-5D-3L index values by 0.03, and a ten-point improvement can lead to an increase by approximately 0.25.

A systematic review of utility values for depression found that the pooled mean (SD) utilities based on studies using the standard gamble as a direct valuation method were 0.69 (0.14) for mild, 0.52 (0.28) for moderate and 0.27 (0.26) for severe major depression. The estimates based on studies using EQ-5D as an indirect valuation method were 0.56 (0.16) for mild, 0.52 (0.28) for moderate and 0.25 (0.15) for severe depression. One recent study regressed PHQ-9 on SF-6D scores, moderate depression 0.56. Our results are largely in line with these aforementioned studies.

There was a consistent difference of about 0.07 EQ-5D scores for the same PHQ-9 score if it represented the baseline or endpoint measurements (figure 1). This is understandable because a patient would rate their health status less satisfactory if they stayed equally symptomatic as before after the treatment and also because it means that they continued to suffer from depression for longer. It is, therefore, reasonable to use the conversion table at baseline for relatively new cases of depression and at end of treatment for more chronic cases (table 1).

An effect size to be typically expected after 2 months of antidepressant pharmacotherapy or psychotherapy over the pill placebo condition is 0.3. Given that the average SD of PHQ-9 in the studies was about 6, an effect size of 0.3 corresponds to a difference by two points on PHQ-9. The differences in EQ-5D scores corresponding with the end-of-treatment PHQ-9 scores of x versus x+2, where x is between 5 and 15 (table 1), ranges between 0.08 and 0.13, producing an approximate average of 0.1 EQ-5D scores. If we assume that the same difference would continue for the ensuing 10 months, the gain in QALY per year would be equal to 0.09 QALY; if we assume that the difference

would eventually wear out over the course of the year due to naturalistic improvements to be expected in the control group, the gain in QALY per year would be equal to 0.05 QALY. (See figure 3 for a schematic drawing to help understand the calculation of QALYs based on changing EQ-5D scores. In reality, the changes will be more smoothly curvilinear but the calculation will be similar.) Since one QALY is typically valued at US$50 000 or 3000 Stirling pounds, such therapies would be cost-effective if they cost US$2500 to US$4500 (150 to 270 pounds) or less. If a 1 day fill of generic selective serotonergic reuptake inhibitor

<table>
<thead>
<tr>
<th>Total scores</th>
<th>Change scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ-9</td>
<td>EQ-5D pretreatment</td>
</tr>
<tr>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>0.91</td>
</tr>
<tr>
<td>7</td>
<td>0.89</td>
</tr>
<tr>
<td>8</td>
<td>0.88</td>
</tr>
<tr>
<td>9</td>
<td>0.85</td>
</tr>
<tr>
<td>10</td>
<td>0.84</td>
</tr>
<tr>
<td>11</td>
<td>0.83</td>
</tr>
<tr>
<td>12</td>
<td>0.80</td>
</tr>
<tr>
<td>13</td>
<td>0.77</td>
</tr>
<tr>
<td>14</td>
<td>0.77</td>
</tr>
<tr>
<td>15</td>
<td>0.73</td>
</tr>
<tr>
<td>16</td>
<td>0.72</td>
</tr>
<tr>
<td>17</td>
<td>0.69</td>
</tr>
<tr>
<td>18</td>
<td>0.69</td>
</tr>
<tr>
<td>19</td>
<td>0.65</td>
</tr>
<tr>
<td>20</td>
<td>0.57</td>
</tr>
<tr>
<td>21</td>
<td>0.41</td>
</tr>
<tr>
<td>22</td>
<td>0.34</td>
</tr>
<tr>
<td>23</td>
<td>0.27</td>
</tr>
<tr>
<td>24</td>
<td>0.23</td>
</tr>
<tr>
<td>25</td>
<td>0.17</td>
</tr>
<tr>
<td>26</td>
<td>0.10</td>
</tr>
<tr>
<td>27</td>
<td>0.04</td>
</tr>
<tr>
<td>28</td>
<td>–</td>
</tr>
<tr>
<td>29</td>
<td>–</td>
</tr>
<tr>
<td>30</td>
<td>–</td>
</tr>
<tr>
<td>31</td>
<td>–</td>
</tr>
<tr>
<td>32</td>
<td>–</td>
</tr>
<tr>
<td>33</td>
<td>–</td>
</tr>
<tr>
<td>34</td>
<td>–</td>
</tr>
<tr>
<td>35</td>
<td>–</td>
</tr>
<tr>
<td>36</td>
<td>–</td>
</tr>
<tr>
<td>37</td>
<td>–</td>
</tr>
<tr>
<td>38</td>
<td>–</td>
</tr>
<tr>
<td>39</td>
<td>–</td>
</tr>
<tr>
<td>40</td>
<td>–</td>
</tr>
<tr>
<td>41</td>
<td>–</td>
</tr>
<tr>
<td>42</td>
<td>–</td>
</tr>
</tbody>
</table>

EQ-5D, Euro-Qol Five Dimensions; PHQ-9, Patient Health Questionnaire-9.
Health economics

Figure 3 A schematic graph showing gains in QALY due to typical pharmacotherapies or psychotherapies. A patient may start with PHQ-9 of 20, corresponding with EQ-5D index value of 0.5. Then they may improve after 2 months of antidepressant therapy to EQ-5D score of 0.9 (solid line), while they may improve to EQ-5D score of 0.8 even if on placebo (dashed line). If we assume that the same difference would continue for the ensuing 10 months while showing slow gradual improvement in both cases, the gain in QALY per year would be equal to 0.09 QALY; if we assume that the difference would eventually wear out over the course of the year due to naturalistic improvements to be expected in the control group, the gain in QALY per year would be equal to 0.05 QALY. Please note that this is a schematic drawing for illustrative purposes: in reality, the changes will be more smoothly curvilinear but the calculation will be similar. EQ-5D,Euro-Qol Five Dimensions; PHQ-9, PatientHealth Questionnaire-9; QALY, quality-adjusted-life years.

antidepressants costs 1–3 dollars and a 1-year prescription costs US$400–1200 dollars, or if 8–16 sessions of psychotherapy cost US$1600–3200 dollars, both therapies would be deemed largely cost-effective. An individual’s decision, by contrast, will and should be more variable and no one can categorically reject nor require such treatments for all patients.

Several caveats should be considered when interpreting the results. First, our sample was limited to participants of trials of iCBT. It may be argued that the results, therefore, would not apply to patients with depression undergoing other therapies or in other settings. Second, the correlations between PHQ-9 and EQ-5D were strong enough for total scores at endpoint and for change scores to justify linking but were somewhat weaker at baseline, probably due to limited variability in PHQ-9 scores at baseline because some studies required minimum depression scores. However, the overall correspondence between PHQ-9 scores and EQ-5D had the same shape between baseline and endpoint, which will increase credibility of the linking at baseline as well. Third, we were able to compare PHQ-9 to EQ-5D-3L only. The EQ-5D-5L, which measures health in five levels instead of three, has been developed to be more sensitive to change and to milder conditions. When data become available, we will need to link PHQ-9 and EQ-5D-5L to examine if we can obtain similar conversion values.

Our study also has several important strengths. First, our sample included patients with subthreshold depression and major depression and from the community or workplace and the primary care. Furthermore, they encompassed mild through severe major depression in approximately equal proportions. Second, all the patients in our sample received iCBT or control interventions including care as usual. Potential side effects of different antidepressants, repetitive brain stimulation, electroconvulsive therapy and other more aggressive therapies must of course be taken into consideration when evaluating their impacts, but our estimates, arguably independent of major side effects, can better inform such considerations. Finally, unlike any prior studies, we were able to link specific PHQ-9 scores and their changes scores to EQ-5D-3L index values.

CONCLUSION AND CLINICAL IMPLICATIONS

In conclusion, we constructed a conversion table linking the EQ-5D, the representative generic preference-based measure of health status, and the PHQ-9, one of the most popular depression severity rating scale, for both its total scores and change scores. The table will enable fine-grained assessment of burden of depression at its various levels of severity and of impacts of its various treatments which may bring various degrees of improvement at the expense of some potential side effects.

Author affiliations

1Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, Japan
2Department of Community Mental Health, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel
3Department of Clinical Psychology and Psychotherapy, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
4Department of Clinical Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
5Department of Health Sciences, University of York, York, UK
6Population Health Sciences & National Institute for Health Research Bristol Biomedical Research Centre, University of Bristol, Bristol, UK
7Primary Health Care, School of Public Health and Community Medicine, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden
8Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK
9Aragon Institute for Health Research (IIS Aragón), Miguel Servet University Hospital, Zaragoza, Spain
10Primary Care Prevention and Health Promotion Research Network, RedIAPP, Madrid, Spain
11Faculty of Medicine, School of Public Health, Imperial College London, London, UK
12School of Sociology & Social Policy and Institute of Mental Health, University of Nottingham, Nottingham, UK

Twitter Toshi A Furukawa @Toshi_FRKW and Stephen Z Levine @szlevine

Contributors TAF and EK conceived the study. TAF and SZL designed the study. PC and EK selected the studies and collected, cleaned and combined the IPD. CBu, DDE, SG, SB, DK, MK, CBj, AK, AvS, HR, JM-J, GC, RP and JS contributed to the IPD. SZL and TAF analysed the data and interpreted the results. TAF wrote the initial draft manuscript, and all authors provided critical input and revisions to the draft manuscript and approved the final manuscript.

Funding This study was supported in part by JSPS Grant-in-Aid for Scientific Research (grant number 17K19808) to TAF. EK was supported by the Netherlands Organisation for Health Research and Development (NWO; project number 019.1825.G001). JM-M is supported by the WellcomeTrust Grant (104908/Z/14/Z).

Disclaimer The views expressed are those of the authors and not necessarily those of any of the funding agencies listed above.

Competing interests TAF reports grants and personal fees from Mitsubishi-Tanabe, personal fees from MSD, personal fees from Shionogi, outside the submitted work; In addition, TAF has a patent 2018-177688 concerning smartphone CBT apps pending, and intellectual properties for Kokoro-app licensed to Tanabe-Mitsubishi. JM-M is supported by the Wellcome Trust Grant (104908/2/14/Z).

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The overall database used for this IPD is restricted due to data sharing agreements with the research institutes where the studies were conducted. IPD from individual studies are available from the individual study authors.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMI Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.
REFERENCES


