Journal of Clinical Epidemiology 67 (2014) 1035–1043

Journal of Clinical Epidemiology

The agreement between proxy and self-completed EQ-5D for care home residents was better for index scores than individual domains☆

Angela Devine[a,*], Stephanie J.C. Taylor[b], Anne Spencer[c], Karla Diaz-Ordaz[a], Sandra Eldridge[b], Martin Underwood[d]

[a]Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, London, United Kingdom WC1H 9SH

[b]Queen Mary University of London, 58 Turner Street, London, United Kingdom, E1 2AB

[c]University of Exeter Medical School, Veysey Building, Salmon Pool Lane, Exeter, United Kingdom, EX2 4SN

[d]Warwick Medical School, Coventry, United Kingdom, CV4 7AL

Accepted 9 April 2014; Published online 15 May 2014

Abstract

Objective: Proxy measures are an alternative source of data for care home residents who are unable to complete the health utility measure, but the agreement levels between residents and care home staff for the EQ-5D have not been investigated previously. The objective of the present study was to examine the inter-rater agreement levels for the reporting of EQ-5D by care home residents and staff, adjusting for the impact of clustering.

Study Design and Setting: The data consist of EQ-5D scores for 565 pairs of care home residents and proxies and quality-adjusted life-years (QALYs) for 248 pairs. Cluster-adjusted agreement was compared for the domains, index scores, and QALYs from the EQ-5D. Factors influencing index score agreement are also described.

Results: The results show poor to fair agreement at the domain level (cluster-adjusted Kappa −0.03 to 0.26) and moderate agreement at the score level (cluster-adjusted intra-class correlation coefficient [ICC] 0.44−0.50) and for QALYs (cluster-adjusted ICC 0.59). A higher likelihood of depression and lower cognitive impairment were both associated with smaller discrepancy between proxy and self-completed scores.

Conclusion: Proxies appear to be an acceptable source of data for index scores and QALYs but may be less reliable if individual domains are considered.

Keywords: EQ-5D; Proxy-rated health; Self-rated health; Health-related quality of life; Care homes; Agreement

1. Introduction

One of the biggest changes for health and social care systems is the increase in the number of people aged 85 years and over needing residential care. In 2010, there were 1,410,700 people aged 85 or older in the United Kingdom [1], an increase from 603,400 in 1981 [2]. For people who are aged 85 or older, just under 16% will be at the risk of being in a care home with or without nursing [3]. For many care home residents, quality of life is poor with nearly 90% requiring care because of disability from long-term conditions [4]. Over one-quarter will die within 9 months of admission with most of these deaths occurring in the care home [5]; however, the evidence base for improving the quality of care in care homes is underdeveloped [6]. There is, thus, a pressing need for robust trials testing the effectiveness and cost-effectiveness of interventions to improve the quality of life for care home residents.

Dementia is now the pre-eminent diagnosis among those admitted to care homes in the UK [7]. Additionally, many care home residents are too frail or cognitively impaired to consistently provide quality of life data. For populations where it is difficult to collect self-reported quality of life, alternative sources, including proxy measures, have been used. A preference-based measure, such as the EQ-5D, has good construct validity for self-report [8] and proxy completion [9] and is the preferred approach in end of life

Funding: Funding for the OPERA clinical trial was provided by National Institute for Health Research Health Technology Assessment Programme (project reference 06/02/01).

* Corresponding author: Tel.: +44-(0)-2079272081; fax: +44-(0)-2079272701.

E-mail address: angela.devine@lshtm.ac.uk (A. Devine).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2014.04.005

What is new?

Key findings

- Despite weaker agreement at the domain level, the index scores and QALYs had better agreement between self-reported and proxy measures by care home staff.

What this adds to what is known?

- This study is the first to examine the use of care home staff as proxies for the EQ-5D and to investigate how results are impacted by clustering. Adjusting for clustering often changed the strength of agreement. Ignoring the fact that the individuals and their proxies are clustered within the home can result in Kappa measures appearing to be in better agreement.

What is the implication, what should change now?

- Our findings suggest that the proxy EQ-5D measure by care home staff is likely to be an acceptable alternative to self-reported EQ-5D for index scores and QALYs but is unlikely to have sufficient agreement at the domain level.
- Further studies of outcome measures in care home residents should use agreement measures that adjust for clustering and could investigate to what extent our findings are replicable in other settings and patient groups.

care, where treatment is about caring for the patient rather than curing disease [10]. Furthermore, it has been used in previous studies of care home residents [11]. Proxy EQ-5D (EQ-5D-P) measures have previously been compared with self-completed EQ-5D (EQ-5D-S) measures at discrete time intervals [12–14] and over time [15]; however, EQ-5D-P has not been assessed in the frail elderly populations living in care homes. Care homes pose additional challenges to the use of proxy measures. Many care home residents have no regular visitors, so care home staff are important potential proxies for this group. A further challenge is that many complex interventions for this population are, of necessity, whole home interventions. This means that many studies will be cluster randomized trials where the unit of randomization is the care home. With proxy completion, the same individual may be providing proxy values for multiple residents, adding to the systematic variation at the home level. This may have further consequences for the interpretation and analysis of proxy measures.

This study is the first to examine the use of care home staff as proxies for the EQ-5D and to investigate how results are impacted by clustering. Clustering occurs when individuals within a cluster are more likely to be similar in their characteristics than those outside the clusters. In this article, we consider the agreement of proxy and self-completed EQ-5D data collected during a large cluster randomized controlled trial in a care home setting in the UK. We also investigate whether proxies have a greater central tendency than residents, choosing ‘some problems’ more frequently at the domain level. Finally, we explore which explanatory variables might influence the differences in level of agreement for index scores in our study.

2. Methods

2.1. Participants and setting

All participants were taking part in the Older People's Exercise intervention in Residential and nursing Accommodation trial, which is described in detail elsewhere [16–18]. Briefly, a cluster randomized design was used to compare a whole home intervention to increase physical activity backed up with a twice weekly physiotherapist led exercise session with a depression awareness training programme for care home staff. The economic evaluation of the trial included 798 residents recruited from 78 care homes. Residents who had severe communication problems or who were seen as too ill to be seen at the time of assessment were excluded from the trial. The primary outcome was depressive symptoms measured using the Geriatric Depression Scale-15 (GDS-15) [19]. For the comparison of EQ-5D-S and EQ-5D-P, data were collected alongside the clinical data at baseline, 6 and 12 months. Matched pairs of data were obtained on 565, 402, and 360 residents, respectively, at each time point. Two hundred forty-eight residents had EQ-5D-S and EQ-5D-P scores at all three time points. Intra-cluster coefficients (ρ) were calculated to see the level of clustering of the EQ-5D-S and EQ-5P-P measures.

The care homes were instructed that the EQ-5D-P should be completed by the carer who was most familiar with the resident, either a care staff member working with the resident on a daily basis, or the care home manager. No further data were collected about the care home staff member who completed the survey. In the UK, care assistants form the vast majority of staff in care homes. In care homes, a registered nurse (often also the manager) must be in charge. Physiotherapists and other types of health care professionals are rarely, if ever, employed by UK care homes.

2.2. Data collection

Care homes were recruited to the study, randomized between December 15, 2008 and April 9, 2010, and the follow-up of residents was completed 1 year later [18]. The EQ-5D is a widely used brief generic measure of health utility [11,20,21] used to measure quality of life across disease groups using questions with a scale of 1 (no problems) to 3 (severe problems) in five domains followed by a Visual Analogue Scale (VAS) of their overall health state. The domain responses were transformed into utilities (index scores) derived from the UK general population [22]. For those residents with EQ-5D-S and EQ-5D-P at all three time points, quality-adjusted life-years (QALYs) were calculated using the area under the curve. Residents completed the EQ-5D by themselves in the presence of a research nurse. Where residents were unable to read or write on the forms themselves (eg, because of visual impairment or difficulty using a pencil) the research nurses read the questionnaire out to the residents and were amanuenses. As the use of the VAS may be difficult for those with substantial cognitive or visual impairment, the research nurses indicated if they felt the resident's response was unreliable (eg, if resident without visual impairment was not actually looking at the instrument when they completed it, the study nurse might consider their response to be unreliable). On the same day that the residents completed the EQ-5D, proxy EQ-5D forms were distributed but proxies may have completed the form at a later, but still proximate, date. Proxies were instructed to consider the proxy-patient perspective with the following statement: ''*The following questions ask how you consider the resident would rate their own general health state at the moment. By placing a cross in one box in each group, please indicate which statement best describes how the resident would rate their health state today.*''

In addition to the EQ-5D, other self-completed measures during the trial included the Geriatric Depression Scale score-15 (GDS-15) [19] and the Mini-Mental State Examination (MMSE) [23]. The GDS-15 is a widely used self-rated measure of presence of depressive symptoms; higher scores indicate significant depression. The MMSE is the most widely used measure of cognitive impairment worldwide; lower scores indicate greater cognitive impairment.

2.3. Analysis

The percentage of residents and proxies choosing each level of each domain was compared with examine whether responders have a consistent preference for one of the levels. The level of inter-rater reliability between the proxy and self-completed responses was examined by calculating Kappa coefficients [24] and percent agreement for each of the five domains of the EQ-5D at baseline, 6 and 12 months where

$$\text{Kappa} = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

We used a weighted Kappa using linear weights [25,26] to distinguish between large and small differences in ratings, whereas assigning equal importance to disagreement between different levels of each domain [27,28]. We obtained 95% confidence intervals by bootstrapping (using the kapci command in STATA, with 1,000 replications) [29]. The Kappa coefficient ranges from −1 to 1, and the following standards for interpreting strength of agreement were used: less than 0.00 as poor, 0.00 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 and above as almost perfect [30]. For EQ-5D index scores, VAS and total QALYs, the level of agreement between proxy and self-completed responses was examined by calculating the intra-class correlation coefficients (ICCs) using one-way random-effects analysis of variance (ANOVA) [31] (Note that we use ICC for intra-class and ρ for intra-cluster coefficients in this article). The same benchmarks as those used for Kappas were chosen for the ICC to add comparability between the agreement found at the domain and index score level. Agreement was regarded as ''satisfactory'' if the strength of agreement is 0.7 or higher [32].

Importantly, the calculation of the Kappa and ICC relies on the assumption of independence of the observations, so we adjusted for clustering. To exemplify the effects of clustering, we present both the adjusted and unadjusted results here. In our study, clustering could arise at the home level and also from the exercise intervention, which has the potential to raise morale throughout the care home, affecting the attitude of both patients and care home staff. We calculated Kappa coefficients and their variances for each cluster. To adjust for clustering, we obtained an estimate of the overall within cluster variance, σ_w^2, based on the average of variances of the Kappas for each cluster. We also estimated between cluster variance σ_b^2, from the Kappas for each cluster. We then constructed a cluster-adjusted kappa, as a weighted mean basing the weights w_i on the minimum variance

$$w_i = \frac{n_i}{(1 + (n_i - 1)\rho_\kappa)}$$

where n_i is the size of cluster i and ρ is the intra-cluster correlation coefficient for the Kappas $\rho_\kappa = \frac{\sigma_b^2}{(\sigma_w^2 + \sigma_b^2)}$.

The variance of the cluster-adjusted Kappa was obtained following Kerry and Bland [33], using

$$\sigma^2_{\text{clustered }\kappa} = \frac{\sum_{i=1}^k w_i^2\left(\sigma_b^2 + \frac{\sigma_w^2}{n_i}\right)}{\left(\sum_{i=1}^k w_i\right)^2}.$$

Using the resulting standard errors, we obtained 95% confidence intervals, based on a normal approximation. Any clusters that had only one person, or where perfect agreement was found for all members of the cluster, or where there was no variability within the cluster (ie, all the pairs proxy self-completed were identical within a cluster) were necessarily dropped from the analysis. To check that this change in population did not affect the unadjusted Kappas, these were re-run using the slightly smaller

Table 1. Characteristics of care home residents

Characteristic	Mean or *N* (%) at baseline
Age (yr) at baseline (n = 564)	86.2 (SD = 7.4)
Female (n = 565)	430 (76%)
Age left full-time education (yr) (n = 499)	15.0 (SD = 1.9)
Length of stay in care home (yr) (n = 561)	2.4 (SD = 2.6)
Dementia in medical record (n = 562)	124 (22%)
MMSE[a] (n = 560)	19.0 (SD = 6.5)
GDS[b] (n = 561)	4.6 (SD = 3.1)
EQ-5D-S intra-cluster coefficient (ρ) (n = 565)	0.19 (SE = 0.04)
EQ-5D-P intra-cluster coefficient (ρ) (n = 565)	0.15 (SE = 0.04)

Abbreviations: EQ-5D-P, proxy EQ-5D; EQ-5D-S, self-completed EQ-5D; GDS-15, Geriatric Depression Scale score-15; MMSE, Mini-Mental State Examination; SD, standard deviation; SE, standard error; ρ, intra-cluster coefficients.

[a] MMSE measures cognitive impairment on a scale of 0 to 30.

[b] GDS indicates a presence or absence of depressive mood on a scale of 0 to 15.

population used for the cluster-adjusted Kappas. Table 3 reports the number of clusters and range in cluster size.

To adjust for clustering in the ICC, nested one-way ANOVAs were calculated using three-level random intercept models (with an extra level for care home). With this analysis, we obtained the proportion of the total variance explained by the scores belonging to the same individual, whereas at the same time acknowledging that the individuals are clustered within homes. Hence, the association between proxy and self-completed scores was measured, adjusting for the clustering of residents within care homes. To supplement the ICC, Bland–Altman plots (graphing the relative differences between the EQ-5D-S and EQ-5D-P measures by the average of the EQ-5D-S and EQ-5D-P measures) were constructed [34]. VAS scores are reported twice: once using all scores and once using only those deemed reliable by the research nurses.

We explore the influence of differences in EQ-5D-S and EQ-5D-P on the resulting QALYs, an important consideration when designing a trial for this population. Finally, a multilevel mixed-effects linear regression was used to consider the extent to which the difference in EQ-5D-S and EQ-5D-P was affected by other variables at baseline. These variables were resident age, resident sex, MMSE, GDS-15 and whether in treatment or control arm of the trial, with a random effect for home (to adjust for clustering).

3. Results

3.1. Participants

Table 1 provides descriptive statistics on care home residents with EQ-5D-S and EQ-5D-P scores at baseline. Most residents (76%) were female with a mean age of 86. Not all of these had both proxy and self-completed VAS scores (Table 2). The ρ values indicate that the levels of clustering of measures within homes were fairly high, and we report cluster-adjusted measures for our results unless otherwise indicated. At 6 months, ρ was 0.16 for EQ-5D-S and 0.18 for EQ-5D-P, and at 12 months ρ was 0.14 and 0.30, respectively.

3.2. EQ-5D data

At baseline, the mean index score and VAS given by the residents (index score = 0.56, VAS = 67.2) were higher than that given by the proxy (index score = 0.51, VAS = 64.4). Scores from 17% (96/565) of the residents and 4% (24/565) of the carers reported perfect health (11111) at baseline, and so were assigned an index score of 1. The distribution of index scores for the residents and carers is broadly similar to those reported by Brazier et al. [35], with few reporting values between 1 and 0.95 and few reporting values around 0.45. The worst imaginable health state (33333) at baseline was rare, with one resident and none of the carers providing this rating. For the baseline VAS, 7% (35/474) of the residents and 1% (3/474) of the proxies rated health as perfect (100). No proxies gave worst possible health for the VAS and only nine (2%)

Table 2. Subject-proxy agreement using percent with exact agreement and Kappa values for EQ-5D domains at three points in time

Domain	Time point	% Exact agreement	Kappa coefficient	Kappa adjusted for clustering	Number of clusters (range in cluster size)[a]
Mobility	Baseline (n = 565)	62	0.42 (0.35–0.48)	0.21 (0.12–0.30)	59 (3–15)
	6 mo (n = 402)	55	0.33 (0.25–0.41)	0.23 (0.13–0.34)	52 (2–14)
	12 mo (n = 360)	61	0.45 (0.37–0.53)	0.34 (0.22–0.46)	50 (2–14)
Self-care	Baseline (n = 565)	45	0.22 (0.17–0.29)	0.12 (0.03–0.20)	66 (2–15)
	6 mo (n = 402)	43	0.20 (0.13–0.27)	0.13 (0.04–0.21)	57 (3–14)
	12 mo (n = 360)	47	0.26 (0.19–0.34)	0.14 (0.06–0.23)	54 (2–14)
Usual activities	Baseline (n = 565)	42	0.09 (0.02–0.16)	0.03 (−0.12, 0.06)	64 (2–15)
	6 mo (n = 402)	52	0.21 (0.13–0.30)	0.21 (0.10–0.31)	49 (3–14)
	12 mo (n = 360)	49	0.19 (0.10–0.28)	0.14 (0.04–0.24)	49 (2–14)
Pain/discomfort	Baseline (n = 565)	58	0.29 (0.22–0.35)	0.21 (0.12–0.30)	59 (3–15)
	6 mo (n = 402)	57	0.25 (0.18–0.33)	0.21 (0.10–0.32)	51 (2–14)
	12 mo (n = 360)	55	0.25 (0.16–0.33)	0.27 (0.18–0.37)	52 (2–14)
Anxiety/depression	Baseline (n = 565)	57	0.23 (0.16–0.31)	0.17 (0.08–0.26)	61 (4–15)
	6 mo (n = 402)	52	0.12 (0.04–0.21)	0.04 (−0.07, 0.14)	45 (3–14)
	12 mo (n = 360)	57	0.14 (0.04–0.23)	0.07 (−0.05, 0.20)	40 (2–14)

[a] For the Kappa adjusted for clustering.

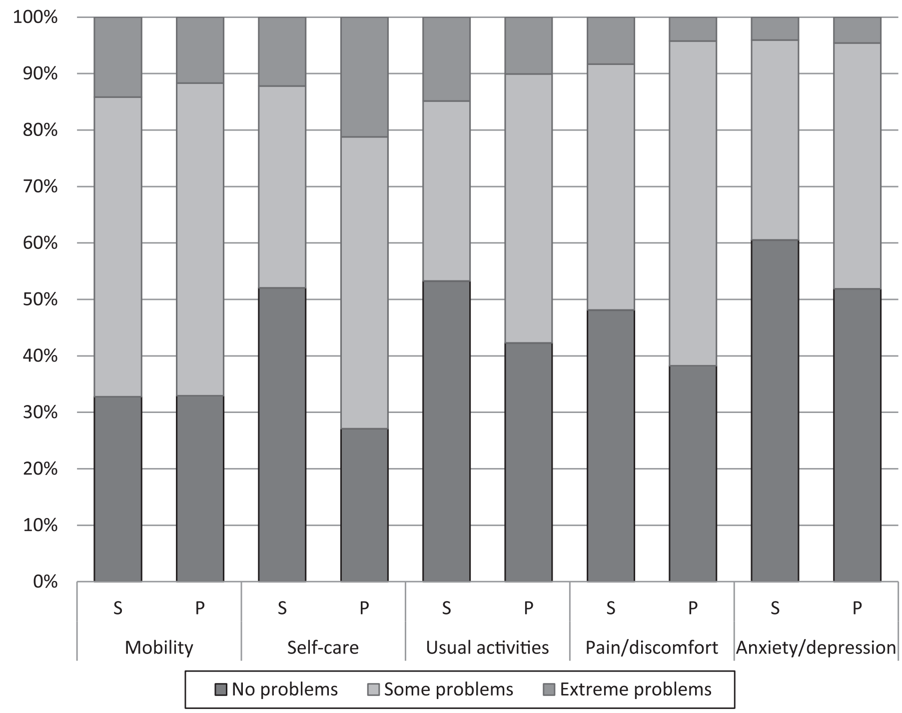

Fig. 1. Distribution of responses to domains of EQ-5D (n = 565). (S = self-completed, P = proxy).

of the residents rated their health as the worst possible. Of these, only three scores were considered reliable by the trial staff. For those with index scores at all three time points, 17% of carers consistently over-rated residents' health and 26% consistently under-rated health status compared with residents' own ratings. Fig. 1 suggests that there is a greater tendency for the proxies to use the some problems category (2) compared with self-report.

3.3. Inter-rater agreement

At the domain level, fair agreement was found between the patient and proxy measures for usual activities (at 6 months) and mobility and pain/discomfort (Table 2). All other domains had slight agreement. When the unadjusted Kappas were re-run using the smaller population used for the cluster-adjusted Kappas, the strength of agreement showed little or no change for each measure.

The ICCs for the VAS indicated slight to fair agreement and moderate agreement for the index score across all time points but stronger agreement for the QALYs; 0.59 regardless of whether adjusted or unadjusted (Table 3). The magnitude for the ICC of the index score was greater than that of the VAS at each time point, lending support to the notion that the VAS has poorer agreement overall. Limiting the ICC to look only at 'reliable' VAS scores (research nurse

Table 3. Mean scores and subject-proxy agreement for QALYs, EQ-5D index scores, and VAS at three points in time

Measure and time	Self-completed mean (SD)	Proxy mean (SD)	ICC (95% CI)	ICC adjusted for clustering (95% CI)
QALYs				
Over 1 yr (n = 248)	0.58 (0.33)	0.55 (0.27)	0.59 (0.50, 0.67)	0.59 (0.50, 0.67)
Index score				
Baseline (n = 565)	0.56 (0.38)	0.51 (0.32)	0.47 (0.40, 0.53)	0.50 (0.40, 0.59)
6 mo (n = 402)	0.57 (0.36)	0.52 (0.32)	0.44 (0.35, 0.51)	0.44 (0.36, 0.52)
12 mo (n = 360)	0.57 (0.38)	0.53 (0.33)	0.50 (0.42, 0.58)	0.50 (0.42, 0.58)
VAS				
Baseline (n = 474)	67.2 (21.2)	64.4 (17.7)	0.16 (0.07, 0.25)	0.16 (0.09, 0.27)
6 mo (n = 342)	67.9 (21.7)	68.2 (18.5)	0.24 (0.14, 0.34)	0.22 (0.14, 0.35)
12 mo (n = 302)	69.7 (21.7)	71.0 (17.8)	0.17 (0.06, 0.28)	0.17 (0.09, 0.32)
VAS: reliable only[a]				
Baseline (n = 425)	67.3 (19.5)	64.4 (17.5)	0.23 (0.14, 0.32)	0.23 (0.15, 0.34)
6 mo (n = 313)	68.0 (20.2)	68.4 (18.7)	0.29 (0.19, 0.39)	0.28 (0.18, 0.40)
12 mo (n = 288)	69.2 (21.5)	71.2 (17.6)	0.18 (0.06, 0.29)	0.17 (0.08, 0.32)

Abbreviations: CI, confidence interval; ICC, intra-class correlation coefficients; QALYs, quality-adjusted life-years; SD, standard deviation; VAS, Visual Analog Scale.

[a] VAS reliable indicates that the research nurse who collected the data thought that the resident understood what was being asked.

Table 4. Results of regressing resident characteristics on the agreement between resident and proxy scores (the difference between EQ-5D-S and EQ-5D-P)

Explanatory variable	Coefficient	SE	*P*-value	95% CI
Age	0.00	<0.01	0.81	<−0.01, <0.01
Male	−0.06	0.03	0.07	−0.12, 0.01
GDS-15	−0.03	<0.01	<0.01	−0.03, −0.02
MMSE	−0.02	<0.01	<0.01	−0.02, −0.01
Intervention arm	−0.02	0.03	0.61	−0.09, 0.05

Abbreviations: CI, confidence interval; EQ-5D-P, proxy EQ-5D; EQ-5D-S, self-completed EQ-5D; GDS, Geriatric Depression Scale score-15; MMSE, Mini-Mental State Examination; SE, standard error.

assessment) increased the agreement, but not to the level of the index score. With a range of 0.44 to 0.50, the ICC for the index score showed negligible fluctuations over time.

Tables 2 and 3 show unadjusted and adjusted (for clustering by care home) measures of agreement. In the cluster-adjusted Kappas, the proportion of variability explained by the relationship between proxy and self-completed responses is reduced, which results in a lowering of the Kappa. For the ICCs (Table 3), the impact of clustering within homes is smaller, and the differences between adjusted and unadjusted measures are smaller.

The Bland−Altman plots did not indicate any important patterns. The difference in QALYs between treatment and control arm in the trial was not significantly different for either EQ-5D-P or EQ-5D-S (data not shown).

3.4. *Influences on agreement*

Results from the mixed-effects model are shown in Table 4. Both GDS-15 and MMSE scores were significantly associated with agreement between EQ-5D-S and EQ-5D-P ($P < 0.05$). A higher GDS-15 score (and higher likelihood of depression) was associated with smaller EQ-5D-S and EQ-5D-P differences whereas a higher MMSE score (and lower cognitive impairment) was also associated with smaller differences. Age, sex, and treatment arm of the trial were not significantly associated with the agreement. It is important to note that with a mean MMSE score of 19 (SD = 6.5), most residents in this study had mild to moderate cognitive impairment. Consequently, the influence of MMSE on the degree of association is only within the range of the MMSE within our sample, and cognition is less likely to have an impact on observed agreement compared with a sample including more severely cognitively impaired patients.

4. Discussion

We believe this is the first study evaluating the reliability of EQ-5D-P in care home residents and to adjust the measures of agreement for clustering. The impact of accounting for the clustering in the data appears to be slightly different between the domain and the index scores, with the cluster-adjusted Kappa of the domains resulting in slightly lower agreement, whereas the impact of adjusting for cluster in the ICCs is virtually negligible. Our study shows fair agreement between the proxy and patient measures for the usual activities domain (at 6 months) and mobility and pain/discomfort domains. Usual activities (at baseline and 12 months), self-care and anxiety/depression showed slight agreement. These findings contrast with Coucill et al. [14] findings of the poorest agreement for the pain/discomfort and anxiety/depression domains. Despite weaker agreement at the domain level, the index score and QALYs had better agreement. We found moderate agreement between EQ-5D-S and EQ-5D-P index scores and slight to fair agreement with the VAS scores. Similar to the results of other proxy studies that examined agreement for both measures, we consistently found better agreement with the index score than the VAS [12,15,36]. Importantly, the total QALYs are not affected by choice of measure. Our result, however, should be interpreted with caution as it reflects a small proportion of the total trial population (31%) and excludes those who died during follow-up.

One explanation for our findings is offered by Kunz who points out that domains have differing impacts on the resulting index score [13]. For example, the usual activities dimension has less impact on the index score than the other domains [22]. Thus, poor to slight agreement between proxy and resident usual activity scores has less bearing on the corresponding index scores, whereas stronger agreement on the pain/discomfort domain has a greater impact on the index score. The stronger agreement between resident and proxy index scores and QALYs lends justification for the use of proxy assessments in clinical trials and appears to reflect other applications, such as the mapping from disease specific to generic EQ-5D states. Previous mapping work has also found a tendency for greater concordance of EQ-5D index scores compared with domain levels [37].

Previous research in the elderly includes a systematic review of health state values for Alzheimer's disease [38]. This review showed that most studies did not systematically collect EQ-5D data from nursing homes residents [39], and some studies excluded patients in nursing homes altogether [15]. A systematic review of proxy EQ-5D measures for dementia patients by Hounsome et al. [40] found only one article that collected proxy data for patients living in institutional settings [12]; however, this research did not focus entirely on the institutional setting and also collected data on patients based in hospital units and at home. Moreover, the percentage of the people living in institutions was not specified, the institutional setting was not described, and nursing staff were the proxies. Our results of moderate agreement for the EQ-5D index score contrast those of the review, which found a general lack of association between self and proxy measures for dementia patients [40]. These differences are likely to reflect the residential and care home population considered here and the smaller proportion of patients with dementia (22%). Furthermore, Hounsome et al.'s use of more stringent benchmarks of agreement than other studies [12,14] is likely to account for some of the observed difference. The percentage of residents reporting perfect health in the EQ-5D index score in our study (17%) was much lower than that found in dementia patients by Coucill et al. (48%) [14] and closer to that found by Kunz (9%) [13] and Ankri et al. (15%) [12]. Unlike Tamim et al.'s study of the elderly after a visit to the emergency department [15], we did not find greater agreement over time, but this is likely to be due to study population differences.

A tendency for proxy values to be lower than self-reported values has been previously reported [13,15] and is also found for other quality of life measures [41]; however, some studies have found this tendency to be reduced in ranking exercises [42]. A tendency for proxy values to be lower than self-report may also explain why depression was associated with smaller differences between EQ-5D-S and EQ-5D-P. Depression affects the residents' overview of other aspects of their health, and so a depressed resident is more likely to view their health as poor, in line with the proxy's assessment. We also found that the differences between EQ-5D-S and EQ-5D-P were smaller with lower cognitive impairment, which has been found elsewhere, and is likely to result in the cognitively impaired being less able to recognize deteriorations in their own health.

Institutional carers are unique proxies in that they spend much more time with the residents than any medical professionals but have a different burden of care to family carers [43]. Carers are not clinically trained, and little is known about their levels of training [44]. In some homes, the EQ-5D-P was completed by the care home manager because of language or literacy barriers or time constraints among the care staff working with the residents on a day-to-day basis. The care home managers may have had a better understanding of the survey but may have had less interaction with the residents which might have affected their interpretation of the residents' quality of life; although, particularly in smaller homes, the managers had a great deal of contact with their residents.

Asking proxies to rate health from the patient's, rather than their own, perspective seeks to align them to the patient's health expectations and to consider the patient's adaptation to symptoms [45]. Consequently, asking proxies to rate health from the patient's perspective rather than their own should be more closely aligned to patient level EQ-5D domain responses. Proxies in our study were asked to consider the patient's perspective, but it is unclear whether there was enough emphasis placed on giving the resident's perspective. Differences between the residents' and proxies' perspectives may only become apparent if they are asked from both perspectives. For example, Pickard et al. asked proxies to complete the EQ-5D from both the patient's and their own perspective and found smaller differences between the self-completed and proxy VAS scores when asked to consider the patient's perspective (3.8 compared with 6.5) [36]. Results of between subject design; however, have more mixed evidence [46,47], indicating that asking the question from both perspectives may be necessary to clarify the perspective that is sought.

This study has some limitations. First, we lack data on which individual proxies completed the surveys at each time point. It is possible that a different proxy completed the survey at each time point as our emphasis was on a proxy who knew the current status of the individual resident very well. Use of institutional carers presents an additional level of clustering to the data because, unlike family members, they are often required to assess multiple patients, but we could not adjust for this without data on the individual proxies. Second, the proxy and resident may have completed the survey on different days. Although our surveys were not designed to allow comparison in dates that the survey was completed, Tamim et al. found that shorter time periods between administering the scales did not lead to greater agreement between patient and proxy measures [15].

From this study, we have a number of recommendations. For studies comparing proxy and self-completed measures, we recommend more transparent reporting of the methods used, which would allow greater comparability across studies. In particular, when reporting Kappa statistics, the type of weighting used and its justification should be described. Similarly, when using ICCs the type of ANOVA used and the justification for that choice should be reported. Given that differences are likely to arise between proxy and self-completion, we recommend that researchers investigate whether these differences translate into differences in overall QALY scores to help justify why one measure is chosen over the other in future studies. We also recommend thinking carefully about proxy survey design and administration to ensure collection of the perspective that is being sought. Against this background, the article also highlights the need to consider carefully the assumption of independence of observations as small changes were seen between ICC measures and larger differences were seen between the Kappa measures. In studies where there is a high degree of clustering of observations, it may be preferable to use self-reported measures to avoid the problems raised by lack of independent observations when the domain-level results are of interest. Further studies should investigate to what extent this finding is replicable in other settings and patient groups. Finally, the frequency of reporting level two, "some problems", could have some bearing upon the number of levels used in the EQ-5D, and research is underway to expand this to a five-level response scale (EQ-5D-5L) [48]. This may simply represent proxies taking the middle ground or representing a degree of uncertainty over the domain akin to the use of 50% when judging risks to reflect uncertainty about the probability of an event [49]. This suggests that a four-level EQ-5D measure might be better than the currently suggested five-level for proxies, but this is beyond the scope of this article [50].

In conclusion, the agreement between the EQ-5D-P and EQ-5D-S improves as it moves from domain to index scores and QALYs, but despite these improvements the level of agreement falls below recommended levels.

Acknowledgments

We would like to acknowledge the helpful comments from Jeff Round, the discussant of our article at the Winter 2013 Health Economists Study Group in Exeter UK (January 9−11, 2013), and for the useful discussion that followed with the other attendees.

M.U., S.J.C.T., A.D., S.E., and A.S. received financial support from the NIHR for work on this study. S.J.C.T., S.E., and M.U. hold other NIHR grants. M.U. receives chairing fees from NICE and has received payment for lectures from BMA learning. S.J.C.T. holds a NIHR HS&RD grant. A.S. also received royalties from Wiley Publishers and is employed by Exeter Medical School. K.D.-O. received no financial support for her efforts on this study.

Department of Health Disclaimer: The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Health Technology Assessment Programme, NIHR, NHS or the Department of Health.

References

[1] Office for National Statistics. Population estimates for UK, England and Wales, Scotland and Northern Ireland, Mid-2010 2011. Available at http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-231847. Accessed September 30, 2013.

[2] Office for National Statistics. Population estimates for UK, England and Wales, Scotland and Northern Ireland, Mid-1981 2003. Available at http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-162562. Accessed September 30, 2013.

[3] Age UK. Later life in the United Kingdom 2014. Available at www.ageuk.org.uk/Documents/EN-GB/Factsheets/Later_Life_UK_factsheet.pdf?dtrk=true. Accessed September 11, 2014.

[4] CCC. Care at the Cross-Road: CCC Care Census Survey Points the Way Forward 2006. Available at www.ccc-ltc.org.uk/pr/2006-07-03.pdf. Accessed January 3, 2012.

[5] Sutcliffe C, Burns A, Challis D, Mozley CG, Cordingley L, Bagley H, et al. Depressed mood, cognitive impairment, and survival in older people admitted to care homes in England. Am J Geriatr Psychiatry 2007;15:708−15.

[6] Tolson D, Rolland Y, Andrieu S, Aquino JP, Beard J, Benetos A, et al. International Association of Gerontology and Geriatrics: a global agenda for clinical research and quality of care in nursing homes. J Am Med Dir Assoc 2011;12:184−9.

[7] Bebbington A, Darton R, Netten A. Care homes for older people: admissions, needs and outcomes. 2001.

[8] McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 3rd ed. Oxford, UK: Oxford University Press; 2006.

[9] Bryan S, Hardyman W, Bentham P, Buckley A, Laight A. Proxy completion of EQ-5D in patients with dementia. Qual Life Res 2005;14:107−18.

[10] Round J. Is a QALY still a QALY at the end of life? J Health Econ 2012;31:521−7.

[11] Borowiak E, Kostka T. Predictors of quality of life in older people living at home and in institutions. Aging Clin Exp Res 2004;16:212−20.

[12] Ankri J, Beaufils B, Novella JL, Morrone I, Guillemin F, Jolly D, et al. Use of the EQ-5D among patients suffering from dementia. J Clin Epidemiol 2003;56:1055−63.

[13] Kunz S. Psychometric properties of the EQ-5D in a study of people with mild to moderate dementia. Qual Life Res 2010;19:425−34.

[14] Coucill W, Bryan S, Bentham P, Buckley A, Laight A. EQ-5D in patients with dementia - an investigation of inter-rater agreement. Med Care 2001;39:760−71.

[15] Tamim H, McCusker J, Dendukuri N. Proxy reporting of quality of life using the EQ-5D. Med Care 2002;40:1186−95.

[16] Underwood M, Eldridge S, Lamb S, Potter R, Sheehan B, Slowther AM, et al. The OPERA trial: protocol for a randomised trial of an exercise intervention for older people in residential and nursing accommodation. Trials 2011;12:27.

[17] Underwood M, Lamb SE, Eldridge S, Sheehan B, Slowther A, Spencer A, et al. Exercise for depression in care home residents: a randomised controlled trial with cost-effectiveness analysis (OPERA). Health Technol Assess 2013;17:1−281.

[18] Underwood M, Lamb SE, Eldridge S, Sheehan B, Slowther AM, Spencer A, et al. Exercise for depression in elderly residents of care homes: a cluster-randomised controlled trial. Lancet 2013;382:41−9.

[19] Yesavage JA, Sheikh JI. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. Clin Gerontol 1986;5:165−73.

[20] Brooks R. EuroQol: the current state of play. Health Policy 1996;37:53−72.

[21] The EuroQol Group. EuroQoL—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199−208.

[22] Dolan P, Gudex C, Kind P, Williams A. A social tariff for Euroqol: results from a UK general population survey. 1995; Discussion Paper: 138, Centre for Health Economics at the University of York.

[23] Folstein MF, Folstein SE, McHugh PR. ''Mini-mental state''. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189−98.

[24] Selai C, Pickard S, Johnson JA. The collection of EuroQol EQ-5D via proxies. EQ Plenary Meeting 2002.

[25] Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. Am J EEG Technol 1971;11:101−10.

[26] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213−20.

[27] Vanbelle S, Albert A. A note on the linearly weighted kappa coefficient for ordinal scales. Stat Methodol 2009;6:157−63.

[28] Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 1996;7:199−202.

[29] Reichenheim ME. Confidence intervals for the kappa statistic. Stata J 2004;4:421−8.

[30] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.

[31] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420−8.

[32] Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193−205.

[33] Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med 2001;20:377−90.

[34] Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465−76.

[35] Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13:873−84.

[36] Pickard AS, Lin HW, Knight SL, Sharifi R, Wu ZG, Hung SY, et al. Proxy assessment of health-related quality of life in African American and White respondents with prostate cancer perspective matters. Med Care 2009;47:176−83.

[37] Hawton A, Green C, Telford CJ, Wright DE, Zajicek JP. The use of multiple sclerosis condition-specific measures to inform health policy decision-making: mapping from the MSWS-12 to the EQ-5D. Mult Scler 2012;18:853−61.

[38] Shearer J, Green C, Ritchie CW, Zajicek JP. Health state values for use in the economic evaluation of treatments for Alzheimer's disease. Drugs Aging 2012;29:31–43.

[39] Naglie G, Tomlinson G, Tansey C, Irvine J, Ritvo P, Black SE, et al. Utility-based quality of life measures in Alzheimer's disease. Qual Life Res 2006;15:631–43.

[40] Hounsome N, Orrell M, Edwards RT. EQ-5D as a quality of life measure in people with dementia and their carers: evidence and key issues. Value Health 2011;14:390–9.

[41] Ferri CV, Pruchno RA. Quality of life in end-stage renal disease patients: differences in patient and spouse perceptions. Aging Ment Health 2009;13:706–14.

[42] Arons AM, Krabbe PF, Scholzel-Dorenbos CJ, van der Wilt GJ, Olde Rikkert MG. Thurstone scaling revealed systematic health-state valuation differences between patients with dementia and proxies. J Clin Epidemiol 2012;65:897–905.

[43] Hung SY, Pickard AS, Witt WP, Lambert BL. Pain and depression in caregivers affected their perception of pain in stroke patients. J Clin Epidemiol 2007;60:963–70.

[44] Bridges J, Brown-Wilson C, Davies S, Dewar B, Dudman J, Heath H, et al. Quality of life in care homes. London, UK: My Home Life; 2006.

[45] Pickard AS, Knight SJ. Proxy evaluation of health-related quality of life: a conceptual framework for understanding multiple proxy perspectives. Med Care 2005;43:493–9.

[46] Lobchuk MM, McClement SE, Daeninck PJ, Shay C, Elands H. Asking the right question of informal caregivers about patient symptom experiences: multiple proxy perspectives and reducing interrater gap. J Pain Symptom Manage 2007;33:130–45.

[47] Gundy CM, Aaronson NK. The influence of proxy perspective on patient-proxy agreement in the evaluation of health-related quality of life: an empirical study. Med Care 2008;46:209–16.

[48] Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.

[49] de Bruin WB, Carman KG. Measuring risk perceptions: whatdoes the excessive use of 50% mean? Med Decis Making 2012;32:232–6.

[50] Choi BCK, Pak AWP. A catalog of biases in questionnaires. Prev Chronic Dis 2005;2:A13. [serial online].