

Proxy reporting of Health-Related Quality of Life using the EuroQol-5D in care home residents; how good is it and how should we deal with clustering effects?

Angela Devine^{*1}, Stephanie J.C. Taylor¹, Anne Spencer², Karla Diaz-Ordaz³, Sandra Eldridge¹, Martin Underwood⁴

¹ Queen Mary University of London; ² University of Exeter Medical School; ³ London School of Hygiene & Tropical Medicine; ⁴ Warwick Medical School

*Corresponding author: angela.devine@lshtm.ac.uk.

ABSTRACT

Background: Many care home residents have marked cognitive impairment and cannot complete the measures of health utility needed for cost-utility analyses. Proxy measures are an alternative source of data when residents cannot provide data directly. The level of agreement in the EQ-5D between care home residents and staff has not been investigated previously, nor has the impact of clustering.

Aims: To identify inter-rater agreement levels for the reporting of EQ-5D by care home residents and carers. To investigate whether proxies have a greater tendency than residents to choose 'some problems' at the domain level. To explore factors influencing agreement of the index score.

Methods: Baseline, 6 and 12 months data were collected from a large cluster randomised clinical trial. This produced 565 sets of paired data at baseline; missing data were not imputed. Weighted Kappas were calculated at the domain level. Intra-class correlation coefficients (ICCs) were calculated for EQ-5D index scores, VAS and total QALYs using one-way random-effects ANOVA. Methods were developed to adjust the Kappa and ICC for clustering. A random effect mixed model was used to investigate other variables influencing observed differences in self-completed and proxy EQ-5D scores at baseline.

Results: We found slight to moderate agreement at the domain level (Kappa 0.122-0.450), slight to fair agreement for the VAS (ICC=0.164-0.242), moderate agreement at the score level (ICC 0.436-0.504) and substantial agreement for the QALYs (ICC 0.629). Proxies had a greater tendency to use the 'some problems' category (2) compared to self-report. The random effect mixed model showed that a higher Geriatric Depression Score-15 (associated with greater likelihood of depression) and lower Mini-Mental State Examination score (associated with greater cognitive impairment) were both associated with more discrepancy between proxy and self-completed scores.

Conclusion: Proxies appear to be a reliable source of data for overall QALY scores but may be less reliable if individual domains are considered. In particular, proxies are less likely to use extreme values. The presence of more severe depression or cognitive impairment in residents seems to reduce their agreement with proxy scores. Given the current movement toward the EQ-5D-5L, the frequency of proxy reporting of 'some problems' in our study suggests that a 4-level survey is worth considering as proxies would not be able to opt for the 'middle' option (i.e. avoids 'central tendency').

INTRODUCTION

One of the biggest challenges for health and social care systems is the current expansion in the number of people aged 85 years and over. There are 18,000-19,000 care homes in England with a capacity of 450,000-469,000 places; the majority of places are residential, not involving specialist nursing care^{1,2}. Nearly 90% of residents in care homes require care because of disability from long-term conditions; 72% have mobility problems; and 62% are described as confused³. Cognitive impairment in the elderly often means that they are unable to provide useable quality of life data, an important patient reported outcome in cost-effectiveness analyses. For other populations where it's difficult to collect self-reported quality of life, alternative sources have been used, including proxy measures. A preference based measure, such as the EQ-5D has good construct validity for self report⁴ as well as proxy completion⁵, but the performance of the EQ-5D measure in frail elderly populations living in care homes had not been compared with proxy measures. One cannot necessarily generalise from studies of free living older people to this environment⁶. Around two thirds of those admitted to care homes have severe cognitive impairment⁷ and over one quarter will die within nine months of admission with the majority of these deaths occurring in the care home⁸. Some care home residents have no regular visitors; therefore, care home staff are important potential proxies for this group but care home proxy responses have not been previously compared to self-reported.

In this paper we consider the agreement of proxy and self-completed EQ-5D data collected during a large cluster randomised controlled trial in a care home setting in the UK. Proxy EQ-5D (EQ-5D-P) measures have previously been compared with self-completed EQ-5D (EQ-5D-S) measures at discrete time intervals⁹⁻¹¹ and over time¹². We focus here upon the use of the EQ-5D measure and its proxy completion as the EQ-5D has been argued to be the preferred approach even in situations, such as end of life care, where treatment is about caring for the patient rather than curing disease¹³. We also consider the impact that clustering might have upon our results. In particular our aims were to:

1. Evaluate the inter-rater agreement between care home residents and care staff EQ-5D domains, index values, visual analogue scale (VAS) scores and QALYs.
2. Investigate whether proxies have a greater central tendency than residents, more frequently choosing 'some problems' at the domain level.
3. Explore which variables might influence the differences in level of agreement for index scores in our study.

METHODS

Participants and setting

All participants were taking part in the OPERA trial, which is described in detail elsewhere¹⁴. Briefly, a cluster randomised 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 GDS-15. 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. 248 residents had EQ-5D-S and EQ-5D-P scores at all three time points. The EQ-5D-P was completed by the carer who was most familiar with the resident, either the care staff working with the resident on a daily basis or the care home manager.

Data collection

The EQ-5D is a widely-used brief generic measure of health utility^{6,15,16} used to measure quality of life across disease groups using questions with a scale of 1 (no problems) to 3 (severe problems) in 5 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 population¹⁷. For those residents

with EQ-5D-S and EQ-5D-P at all 3 time points, quality-adjusted life-years (QALYs) were calculated using the area under the curve. The EQ-5D has been used in previous studies of care home residents⁶, but concerns over missing self-completed data led to the use of the proxy EQ-5D in our study. Residents completed the EQ-5D by themselves in the presence of a trained study research nurse. Where residents were unable to read or write on the forms themselves (e.g. because of visual impairment or difficulty holding a pencil) the research nurses read the questionnaire out to the residents and recorded the responses. As the use of the VAS may be difficult for those with substantial cognitive impairment or visual impairment, the research nurses indicated if they felt the resident's response was unreliable (e.g. 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, proxies were given proxy EQ-5D forms to complete. (These were collected at a later date and may not all have been completed on the same day as the residents' EQ-5Ds). 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."

Other measures collected in addition to the EQ-5D during the trial included the Geriatric Depression Scale-15 (GDS-15)¹⁸, the Mini-Mental State Examination (MMSE)¹⁹ and the Barthel index²⁰. 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. The Barthel index is the most widely used measure of physical disability among older people; lower scores indicate greater disability.

Analysis

The percentage of residents and proxies choosing each level of each domain was compared to examine whether responders have a consistent preference for one of

the levels. The level of agreement between the proxy and self-completed responses was examined by calculating Kappa coefficients²¹ and percent agreement and for each of the five domains of the EQ-5D at baseline, 6 and 12 months. The Kappa coefficient was calculated using the formula:

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

Weighted Kappa²² using linear weights²³ was used to distinguish between large and small differences in ratings while assigning equal importance to disagreement between levels^{24,25}. The Kappa coefficient ranges from -1 to 1, and the following standards for strength of agreement were used: less than 0.00 is poor, 0.00 to 0.20 is slight, 0.21 to 0.40 is fair, 0.41 to 0.60 is moderate, 0.61 to 0.80 is substantial, and 0.81 and above is almost perfect²⁶. Intraclass correlation coefficients (ICCs) were calculated for EQ-5D index scores, VAS and total QALYs using one-way random-effects ANOVA²⁷. Different guidelines have been suggested for interpreting the ICC values with slightly different benchmarks for agreement. The same benchmarks as those used for Kappas were chosen to add comparability between the agreement found at the domain and index score level.

Importantly, the calculation of the Kappa and ICC relies on the assumption of independence of the observations, so the impact that clustering might have upon our results was examined. In our study, clustering may arise from the care manager completing the questionnaire for multiple residents in small homes. Clustering could also arise from the exercise intervention, which has the potential to raise morale throughout the care home, affecting the attitude of care home staff and patients alike. To obtain Kappa coefficients which were adjusted for clustering, we calculated the usual Kappa within each cluster and then average these coefficients weighted by the size of the cluster, similar to a meta-analysis²⁸. We did this using a fixed effects model, therefore assuming that the variability is due to random variation. 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). Within this analysis, we

manually calculated the proportion of the total variance explained by the scores belonging to the same individual. 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 constructed²⁹. VAS scores are reported twice: once using all scores and once using only those deemed reliable by the research assistants.

The influence of differences in EQ-5D-P and EQ-5D-S on the resulting QALYs was also explored, an important consideration when designing a trial for this population. We also checked whether our trial results would have been any different had we used the EQ-5D-P or EQ-5D-S,

Finally, a mixed model was used to consider the extent to which the difference in EQ-5D-P and EQ-5D-S were affected by other variables at baseline. These variables were resident age, resident sex, MMSE, GDS and whether in treatment or control arm of the trial, with a random effect for home (to adjust for clustering).

RESULTS

Participants

Table 1 provides descriptive statistics on care home residents with EQ-5D-P and EQ-5D-S scores at baseline. Most (86%) of the residents were female with a mean age of 86. Not all of these had both proxy and self-completed VAS scores (Table 2).

EQ-5D data

At baseline, the mean index score and VAS given by the residents (index score=0.559, VAS=67.3) was higher than that given by the proxy (index score=0.507, VAS=64.4). Seventeen percent (96/565) of the residents and 4% (24/565) of the carers rated the resident as having perfect health (11111) at

Table 1. Characteristics of care home residents.

Characteristic	Mean (SD) or N(%) at baseline (n=565)
Age (years) at baseline	86.2 (7.4)
Female	430 (76%)
Age left full-time education (years)	15.0 (1.9)
Length of stay in care home (years)	2.4 (2.6)
MMSE ^a	19.0 (6.5)
GDS ^b	4.6 (3.1)
Barthel ^c	59.9 (26.2)

^aMMSE measures cognitive impairment on a scale of 0 to 30. ^bGDS indicates a presence or absence of depressive mood on a scale of 0 to 15. ^cThe Barthel index is a widely used measure of activities of daily living on a scale of 0 to 100²⁰.

baseline, and so were assigned an index score of 1. The distribution of index scores for the residents and carers are broadly similar to those reported by Brazier *et al.* (2004)³⁰, 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 of the residents and none of the carers rating their health as such. For the baseline VAS, 7.4% (35/474) of the residents 0.6% (3/474) of the proxies rated health as perfect (100). No proxies gave worst possible health for the VAS and only 9 (1.9%) of the residents rated their health as the worst possible. Of these, only 3 scores were considered reliable by the trial staff. For those with index scores at all 3 time points, 17% of carers consistently over-rated residents' health and 26% consistently under-rated health status compared with residents' own ratings. Figure 1 suggests that there is a greater tendency for the proxies to use the some problems category (2) compared to self-report.

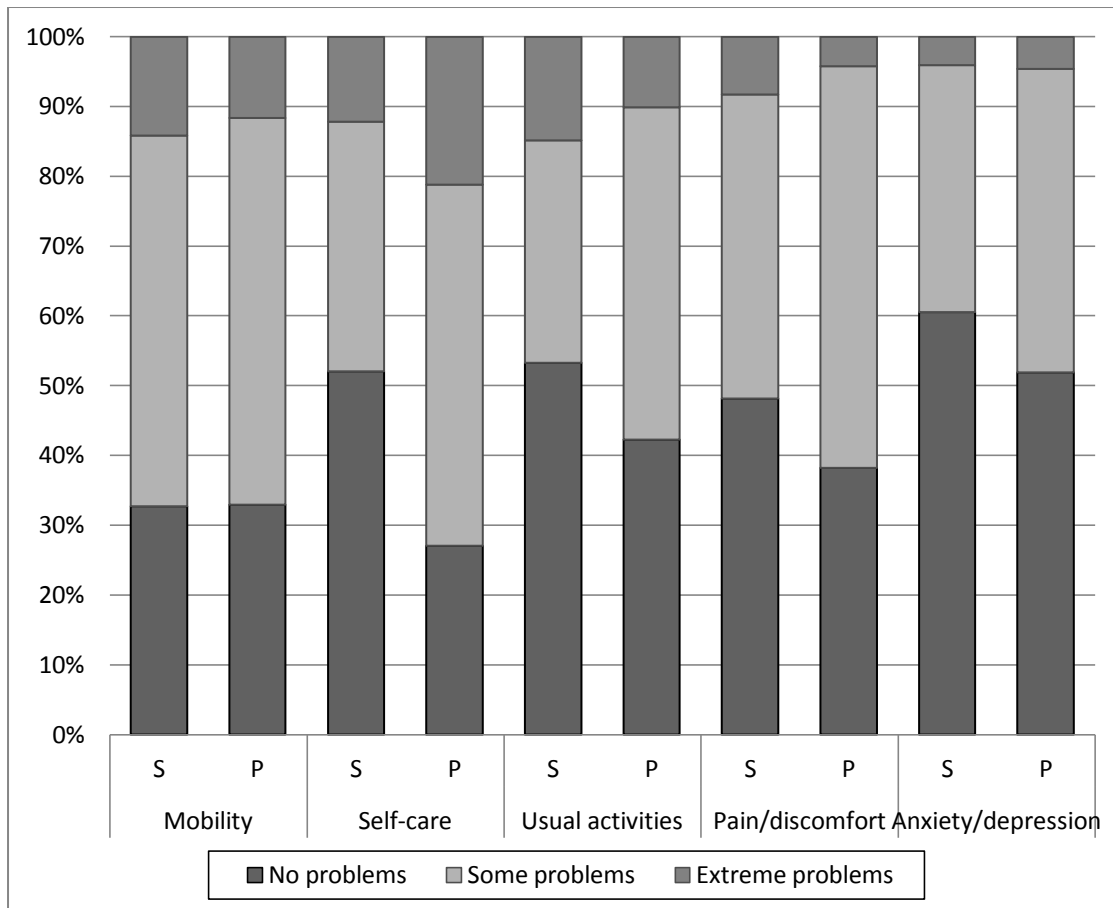


Figure 1. Distribution of responses to domains of EQ-5D (n=565). (S=Self-completed, P=Proxy)

Inter-rater agreement

At the domain level, moderate agreement was seen between the proxy and patient measures for the domains of mobility (at baseline and 12 months), fair agreement for pain and discomfort, but slight to fair on the domains of self-care, usual activities and anxiety and depression (table 2). For each domain, agreement was similar at all 3 time points, indicating that the level of agreement did not change over time. Only the unadjusted Kappas are reported here as the results for Kappas adjusted for clustering are not yet available.

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

Domain	Timepoint	% exact agreement	Kappa coefficient
Mobility	Baseline (n=565)	62	0.418
	6 months (n=402)	55	0.327
	12 months (n=360)	61	0.450
Self-Care	Baseline (n=565)	45	0.222
	6 months (n=402)	43	0.196
	12 months (n=360)	47	0.264
Usual Activities	Baseline (n=565)	42	0.187
	6 months (n=402)	52	0.205
	12 months (n=360)	49	0.187
Pain/Discomfort	Baseline (n=565)	58	0.288
	6 months (n=402)	57	0.254
	12 months (n=360)	55	0.247
Anxiety/Depression	Baseline (n=565)	57	0.230
	6 months (n=402)	52	0.122
	12 months (n=360)	57	0.135

The ICC adjusted for clustering varied between 0.299 and 0.313 for the index score and 0.052 and 0.104 for the VAS (Table 3). The ICC unadjusted for clustering varied between 0.436 and 0.504 for the index score and 0.164 and 0.242 for the VAS. 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 increased the ICC, but not to the level of the index score. The ICC for the index score went down at 6 months and then up at 12 months, indicating that the level of agreement did not change over time.

Table 3 shows that adjusting for clustering consistently lowers the ICC. The adjusted ICCs are lower because they account for the similarities between individuals in the same cluster. Cluster variation is accounted for in these adjusted ICCs, reducing the proportion of variability that is explained by the relationship between proxy and self-completed responses, which results in a lowering of the ICC.

Table 3. Mean scores and subject-proxy agreement for QALYs, EQ-5D index scores and VAS (all and only those with self-completed VAS designated as reliable by the research assistant) at three points in time, and for changes over time.

Measure and time		Self-completed mean (SD)	Proxy mean (SD)	ICC	ICC adjusted for clustering
QALYs	Over 1 year (n=248)	0.511 (0.335)	0.552 (0.271)	0.629	0.391
Index Score	Baseline (n=565)	0.559 (0.377)	0.507 (0.320)	0.467	0.313
	6 months (n=402)	0.573 (0.362)	0.520 (0.323)	0.436	0.299
	12 months (n=360)	0.565 (0.383)	0.530 (0.333)	0.504	0.319
VAS	Baseline (n=474)	67.3 (19.5)	64.4 (17.5)	0.164	0.104
	6 months (n=342)	67.9 (21.7)	68.2 (18.5)	0.242	0.071
	12 months (n=302)	69.7 (21.7)	71.0 (17.8)	0.174	0.052
VAS: reliable only ^a	Baseline (n=425)	67.3 (19.5)	64.4 (17.5)	0.233	0.172
	6 months (n=313)	68.0 (20.2)	68.4 (18.7)	0.291	0.091
	12 months (n=288)	69.2 (21.5)	71.2 (17.6)	0.176	0.050

^aVAS reliable indicates that the research nurse who collected the data thought that the resident understood what was being asked.

The Bland-Altman plots gave no additional information. The difference in QALYs between treatment and control arm in the trial was not significantly different, regardless of whether EQ-5D-P or EQ-5D-S was chosen (data not shown).

Influences on agreement

Results from the mixed model are shown in Table 4. A higher GDS score (associated with greater likelihood of depression) was associated with greater EQ-5D-S and EQ-5D-P differences whilst a higher MMSE score (associated with lower cognitive impairment) was associated with smaller differences. Both GDS and MMSE scores were significantly associated with agreement between EQ-5D-S and EQ-5D-P ($P < 0.05$). Age, sex and treatment arm of the trial were not significantly associated with the agreement.

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	-0.000	0.002	0.807	-0.004 to 0.003
Male	-0.059	0.033	0.074	-0.123 to 0.006
GDS	-0.026	0.004	<0.001	-0.034 to -0.017
MMSE	-0.018	0.002	<0.001	-0.023 to -0.014
Intervention arm	-0.018	0.035	0.613	-0.086 to 0.050
constant	0.580	0.178	<0.001	0.231 to 0.930

DISCUSSION

We believe this is the first study evaluating the performance of EQ-5D-P in care home residents. A systematic review of health state values for Alzheimer's disease only found 1 paper that included data on EQ-5D for those residing in nursing homes³¹, and this study only had only one patient living in a nursing home with the principal carer as the proxy³². Some studies excluded patients in nursing homes altogether¹². A systematic review of proxy EQ-5D measures for dementia patients³³ found only 1 paper that looked at the collection of proxy data by institutional carers. This study by Ankri *et al.* looked at the use of proxy EQ-5D for French patients who were in acute hospital, institutions or dwelling at home³⁴. The percentage of people living in institutions was not specified and the institutional setting was not described. Other differences included Ankri *et al.* using health care professionals as proxies and the delivery of the EQ-5D-S by an un-trained interviewer.

Our study shows moderate agreement between the proxy and patient measures for the mobility domain (at baseline and 12 months), fair for the pain/discomfort domain, and slight to fair on the domains of self-care, usual activities and anxiety/depression. We anticipate that these levels of agreement will drop when the Kappas adjusted for clustering are calculated. Coucill *et al.* found the pain/discomfort and anxiety/depression domains to have the poorest agreement¹¹. It is possible that the anxiety/depression domain may be hard for proxies to imagine. Furthermore, if depression is not recognised by the carer, then this could affect carers' scores and therefore, the level of agreement between proxy and resident. For the usual activities domain, the rating is reliant on subjective assessments that may lead to differing

opinions between the carer and resident. It is less clear why there is lower agreement for self-care dimension. The lower agreement might reflect the residents' high levels of cognitive impairment or arise because the use of commodes and bed washes have become normalised procedures, and so are not perceived by the patient as 'problems with self-care'³⁵.

For the unadjusted ICCs, we found moderate agreement between EQ-5D-S and EQ-5D-P index scores and fair agreement with the VAS scores, similar to the results of other proxy studies that examined agreement for both measures^{9,12}. The percentage of residents reporting perfect health in the EQ-5D index score (17%) was much lower than that found in dementia patients by Coucill *et al.* (48%)¹¹ but higher than that found by Kunz (9%)¹⁰ and Ankri *et al.* (15%)⁹. Unlike Tamim *et al.*'s study of the elderly following a visit to the emergency department¹², we did not find higher 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^{10,36} and also found more broadly for other quality of life measures³⁷.

Our results initially seem at odds with Hounsome *et al.* who concluded that in general there is a lack of association between self and proxy measures of EQ-5D³³. Different benchmarks were used, accounting for some of the differences. Hounsome regarded Kappa coefficients that were less than 0.5 as evidence of a weak association, whereas the benchmarks used by the papers they cite^{9,38} have kappa 0.41 to 0.60 as evidence of moderate agreement²⁶. More importantly, it may be that despite weaker agreement at the domain level, the index score has better agreement; and the total QALYs are not affected. Pickard *et al.* has suggested a greater agreement for the index score³⁹.

Comparing the unadjusted ICCs for Index scores with the Kappa coefficients for the domains, the moderate agreement of the index scores compared favourably to the moderate to poor agreement at the domain level, and the agreement for QALYs was higher still. This indicates that while disagreement may exist at the domain and score level, the overall QALY outcome remained the same. This result, however, should be

interpreted with caution as it reflects a small proportion of the total trial population (31%) and does not include those who died during follow up. One explanation for this finding is offered by Kunz who points out that domains have differing impacts on the resulting index score¹⁰. With the UK tariff¹⁷, the usual activities dimension has less impact on the index score than the other domains, which means that our finding of slight agreement between proxy and resident scores has less bearing on the corresponding index scores. Conversely, differences on the pain/discomfort domain have a greater impact on the index score, but we found fair agreement for this domain. This greater agreement between resident and proxy index scores and QALYs lends justification for the use of proxy assessments in 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 at the EQ-5D population level compared to the domain level⁴⁰.

Institutional carers are unique proxies in that they spend more time with the residents than other medical professionals but have a different burden of care than family carers. Carers are not clinically trained and little is known about their overall levels of training⁴¹. In some homes, the EQ-5D-P was completed by the care home manager due to language or literacy barriers or time constraints amongst the care staff working with the residents on a day to day basis. While these carers may have found it more difficult to complete the survey, they may have had a more comprehensive understanding of what life entailed for the residents. The care home managers, paradoxically, 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.

Proxies in our study were asked to consider the patient's perspective, but it is unclear whether there was enough emphasis placed on the perspective or whether this nuance was understood by the care home staff. Differences between the residents' and proxies' perspectives may only become apparent if they are asked from both perspectives. Demonstrating the importance of distinguishing between the

two perspectives, a study by Pickard *et al.* asked proxies to complete the EQ-5D from both the patient's and their own perspective. They compared VAS scores (but not index scores) and found smaller differences between the self-completed and proxy VAS scores when asked to consider the patient's perspective (3.8 compared to 6.5)³⁹. Results of between subject design, however, have more mixed results with Lobchuk *et al.* recommending the need to emphasize the patient's perspective⁴², while Gunny and Aaronson's research failed to find a compelling case for using either perspective⁴³. Until more research is done in this area, asking the question from both perspectives may help to clarify the perspective that is required and may provide an understanding of a patient's condition; but it is clear that the measures cannot easily be combined⁴⁴.

There are limitations to the study. First, we lack data on who completed the surveys at each time point. Given that managers are likely to have filled out the survey for multiple residents, and there may also be a home effect from our intervention, we have tried to take this into account by adjusting for clustering by home in the analysis. While it is possible that a different proxy completed the survey each time, a low turnover of managers makes this unlikely. Second, the proxy and resident may have completed the survey on different days. While 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 proxy¹².

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.

Our study may also have some bearing upon the number of levels used in the EQ-5D. The EQ-5D measure currently scores each domain using a 3 level response scale but research is underway to expand this to a five level response scale (EQ-5D-5L)⁴⁵. The frequency of reporting level 2, some problems, in our study by proxies, raises some concerns with these developments, particularly when proxy assessment might be used. Level 2 responses 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 reflecting uncertainty about the probability of an event (e.g. 'the central tendency')⁴⁶. This suggests that a 4 level EQ-5D measure might be better than the currently suggested 5 level⁴⁷.

The comparison of proxy and self-completed EQ-5D undertaken here seems to justify the use of proxy measures in this population. Fewer differences were seen at the score and QALY level than domain level, and higher completion rates were observed for proxy completion (72% vs 97% at baseline in the OPERA study). Against this background, the paper also highlights the need to consider carefully the assumption of independence of observations. In studies where observations are clustered, for example, if the proxy completes the EQ-5D on behalf of more than one resident, adjusting for clustering of data points will lower the overall levels of agreement between the proxy and self completed EQ-5D. When there is high degree of clustering (as measured by the intra cluster correlation coefficient), this could justify the use of self-report to avoid the problems raised by the lack of independent of observations. Further studies should investigate to what extent this finding is replicable in other settings and patient groups. In conclusion, the EQ-5D-P is likely to be an acceptable alternative to the EQ-5D-S for the elderly population found in care homes at the QALY level, but is unlikely to have sufficient agreement at the domain level.

REFERENCES

1. The Care Quality Commission. *The state of health care and adult social care in England an- an overview of key themes in care in 2010/11*. London 2011.
2. Office Of Fair Trading. *Care homes for older people in the UK: A market study*. 2005.
3. CCC. Care at the Cross-Road: CCC Care Census Survey Points the Way Forward. 2006; www.ccc-ltc.org.uk/pr/2006-07-03.pdf. Accessed 3rd January, 2012.
4. McDowell I, Newell C. *Measuring health: A guide to rating scales and questionnaires*. 3rd ed. Oxford: Oxford University Press; 2006.
5. Bryan S, Hardyman W, Bentham P, Buckley A, Laight A. Proxy completion of EQ-5D in patients with dementia. *Qual Life Res*. Feb 2005;14(1):107-118.
6. Borowiak E, Kostka T. Predictors of quality of life in older people living at home and in institutions. *Aging Clin. Exp. Res*. Jun 2004;16(3):212-220.
7. Sutcliffe C, Cordingley L, Burns A, et al. A new version of the geriatric depression scale for nursing and residential home populations: the geriatric depression scale (residential) (GDS-12R). *Int Psychogeriatr*. 2000;12(2):173-181.
8. Sutcliffe C, Burns A, Challis D, et al. Depressed mood, cognitive impairment, and survival in older people admitted to care homes in England. *Am J Geriatr Psychiatry*. 2007;15(8):708-715. Epub 2007 May 2015.
9. Ankri J, Beaufile B, Novella JL, et al. Use of the EQ-5D among patients suffering from dementia. *J Clin Epidemiol*. 2003;56(11):1055-1063.
10. Kunz S. Psychometric properties of the EQ-5D in a study of people with mild to moderate dementia. *Quality of Life Research*. Apr 2010;19(3):425-434.
11. Coucill W, Bryan S, Bentham P, Buckley A, Laight A. EQ-5D in patients with dementia - An investigation of inter-rater agreement. *Med. Care*. Aug 2001;39(8):760-771.
12. Tamim H, McCusker J, Dendukuri N. Proxy reporting of quality of life using the EQ-5D. *Med. Care*. Dec 2002;40(12):1186-1195.
13. Round J. Is a QALY still a QALY at the end of life? *Journal of Health Economics*. 2012 May (Epub 2012 Apr 2012;31(3):521-527.
14. Underwood M, Eldridge S, Lamb S, 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.
15. Brooks R. EuroQol: the current state of play. *Health Policy*. 1996;37(1):53-72.
16. The EuroQol Group. A new facility for the measurement of health-related quality of life. 1990:199-208.
17. Dolan P, Gudex C, Kind P, Williams A. A social tariff for Euroqol: results from a UK general population survey. York: Centre for Health Economics; 1995.
18. Yesavage JA, Sheikh JI. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. *Clinical Gerontologist*. 1986/11/18 1986;5(1-2):165-173.
19. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*. Nov 1975;12(3):189-198.
20. Mahoney FI, Barthel DW. FUNCTIONAL EVALUATION: THE BARTHEL INDEX. *Md State Med J*. 1965;14:61-65.
21. Selai C, Pickard S, Johnson JA. The collection of EuroQol EQ-5D via proxies. *EQ Plenary Meeting*. York 2002.
22. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*. Oct 1968;70(4):213-220.
23. Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*. Sep 1971 1971;11(3):101-110.

24. Vanbelle S, Albert A. A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*. 2009;6(2):157-163.
25. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*. 1996;7(2):199-202.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
27. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.
28. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. *Psychological Methods*. 1998 1998;3(4):486-504.
29. Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. *Stat Med*. 1994;13(23-24):2465-2476.
30. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ*. 2004;13(9):873-884.
31. 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(1):31-43.
32. Naglie G, Tomlinson G, Tansey C, et al. Utility-based quality of life measures in Alzheimer's disease. *Quality of Life Research*. May 2006;15(4):631-643.
33. 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 in Health*. Mar-Apr 2011;14(2):390-399.
34. Ankri J, Beaufile B, Novella JL, et al. Use of the EQ-5D among patients suffering from dementia. *J. Clin. Epidemiol*. Nov 2003;56(11):1055-1063.
35. Kutner JS, Bryant LL, Beaty BL, Fairclough DL. Symptom distress and quality-of-life assessment at the end of life: The role of proxy response. *J. Pain Symptom Manage*. Oct 2006;32(4):300-310.
36. Tamim H, McCusker J, Dendukuri N. Proxy reporting of quality of life using the EQ-5D. *Med Care*. 2002;40(12):1186-1195.
37. Ferri CV, Pruchno RA. Quality of life in end-stage renal disease patients: Differences in patient and spouse perceptions. *Aging Ment. Health*. 2009;13(5):706-714.
38. 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(8):760-771.
39. Pickard AS, Lin HW, Knight SL, et al. Proxy Assessment of Health-Related Quality of Life in African American and White Respondents With Prostate Cancer Perspective Matters. *Med. Care*. Feb 2009;47(2):176-183.
40. 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. J*. Jun 2012;18(6):853-861.
41. Bridges J, Brown-Wilson C, Davies S, et al. *Quality of life in care homes*. London: My Home Life;2006.
42. 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*. Feb 2007;33(2):130-145.
43. 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*. Feb 2008;46(2):209-216.
44. Bosboom PR, Alfonso H, Eaton J, Almeida OP. Quality of life in Alzheimer's disease: different factors associated with complementary ratings by patients and family carers. *Int. Psychogeriatr*. May 2012;24(5):708-721.

45. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res.* Dec 2011;20(10):1727-1736. Epub 2011 Apr 1729.
46. de Bruin WB, Carman KG. Measuring Risk Perceptions: What Does the Excessive Use of 50% Mean? *Med. Decis. Mak.* Mar-Apr 2012;32(2):232-236.
47. Choi BCK, Pak AWP. A catalog of biases in questionnaires. *Prev Chronic Dis [serial online]*. 2005.