BMJ Open Quality

Comparison of staff and resident health status ratings in care homes

Tim Benson ,[1,2] Clive Bowman[3]

To cite: Benson T, Bowman C. Comparison of staff and resident health status ratings in care homes. *BMJ Open Quality* 2020;**9**:e000801. doi:10.1136/bmjoq-2019-000801

Received 14 August 2019
Revised 17 December 2019
Accepted 18 February 2020

ABSTRACT

Background Many care home residents cannot self-report their own health status. Previous studies have shown differences between staff and resident ratings. In 2012, we collected 10 168 pairs of health status ratings using the *howRu* health status measure. This paper examines differences between staff and resident ratings.

Method *HowRu* is a short generic person-reported outcome measure with four items: pain or discomfort (discomfort), feeling low or worried (distress), limited in what you can do (disability) and require help from others (dependence). A summary score (*howRu* score) is also calculated. Mean scores are shown on a 0–100 scale. High scores are better than low scores. Differences between resident and staff reports (bias) were analysed at the item and summary level by comparing distributions, analysing correlations and a modification of the Bland-Altman method.

Results and conclusions Distributions are similar superficially but differ statistically. Spearman correlations are between 0.55 and 0.67. For items, more than 92.9% of paired responses are within one class; for the *howRu* summary score, 66% are within one class. Mean differences (resident score minus staff score) on 0–100 scale are pain and discomfort (–1.11), distress (0.67), discomfort (1.56), dependence (3.92) and *howRu* summary score (1.26). The variation is not the same for different severities. At higher levels of pain and discomfort, staff rated their discomfort and distress as better than residents. On the other hand, staff rated disability and dependence as worse than did residents. This probably reflects differences in perspectives. Red amber green (RAG) thresholds of 10 and 5 points are suggested for monitoring changes in care home mean scores.

BACKGROUND

The role of care homes is to provide care and community for a population of people defined by various combinations of mental and physical dependency. Clearly, care home effectiveness should be monitored from a resident perspective whereas typically it is presumed to be good if various processes and regulatory standards are met. Using a simple person-reported outcome measure (PROM) may provide valuable insight but, with typically over 70% of residents having significant cognitive impairment, frailty or being in terminal decline,[1] acquiring survey data is challenging.

An alternative is to ask the staff familiar with the residents to rate them as a proxy. Previous studies have shown that paired assessments by staff and by residents and by staff and relatives give varied results.[2–6] These studies have been small, with varying levels of dementia and did not examine the differences in detail.

The aim of this paper is to assess how well staff and residents agree about perceptions of health status, based on a large sample of paired assessments by staff proxies and residents.[7]

METHOD

Data were collected as part of the 2012 Bupa census, which reported on 24 506 residents in 395 care homes in UK, Australia and New Zealand.

This paper covers 10 168 matched assessments of health status by staff and residents using the *howRu* health status measure.[8] This is a companion to our previous paper, which examined the construct validity of using *howRu*, rated by staff proxies in care homes, to assess resident health status.[9] Full details of the data collection method using optically mark readable forms are provided in that paper.

HowRu is a short generic measure of health-related quality of life or health status. It forms part of a large family of PROMs and person-reported experience measures, completed by patients (or care home residents) and by staff.[10] *HowRu* has been validated for use at the individual patient level,[11] and for construct validity in ambulatory care in comparison with EQ-5D,[12 13] and SF-12.[8]

Resident assessments were collected at the same time as staff assessments and shared the same bar-code identifier. The resident form is shown in figure 1. It also includes a measure of resident experience (*howRwe*),[14] and a version of the Net Promoter Score,[15] but these are not discussed further here.

HowRu asks the question *How are you today?* referring to the past 24 hours; this is the question answered by residents. The question

[1]R-Outcomes Ltd, Thatcham, UK
[2]Institute of Health Informatics, UCL, London, UK
[3]Health Sciences, City University School of Health Sciences, London, UK

Correspondence to
Dr Tim Benson;
tim.benson@r-outcomes.com

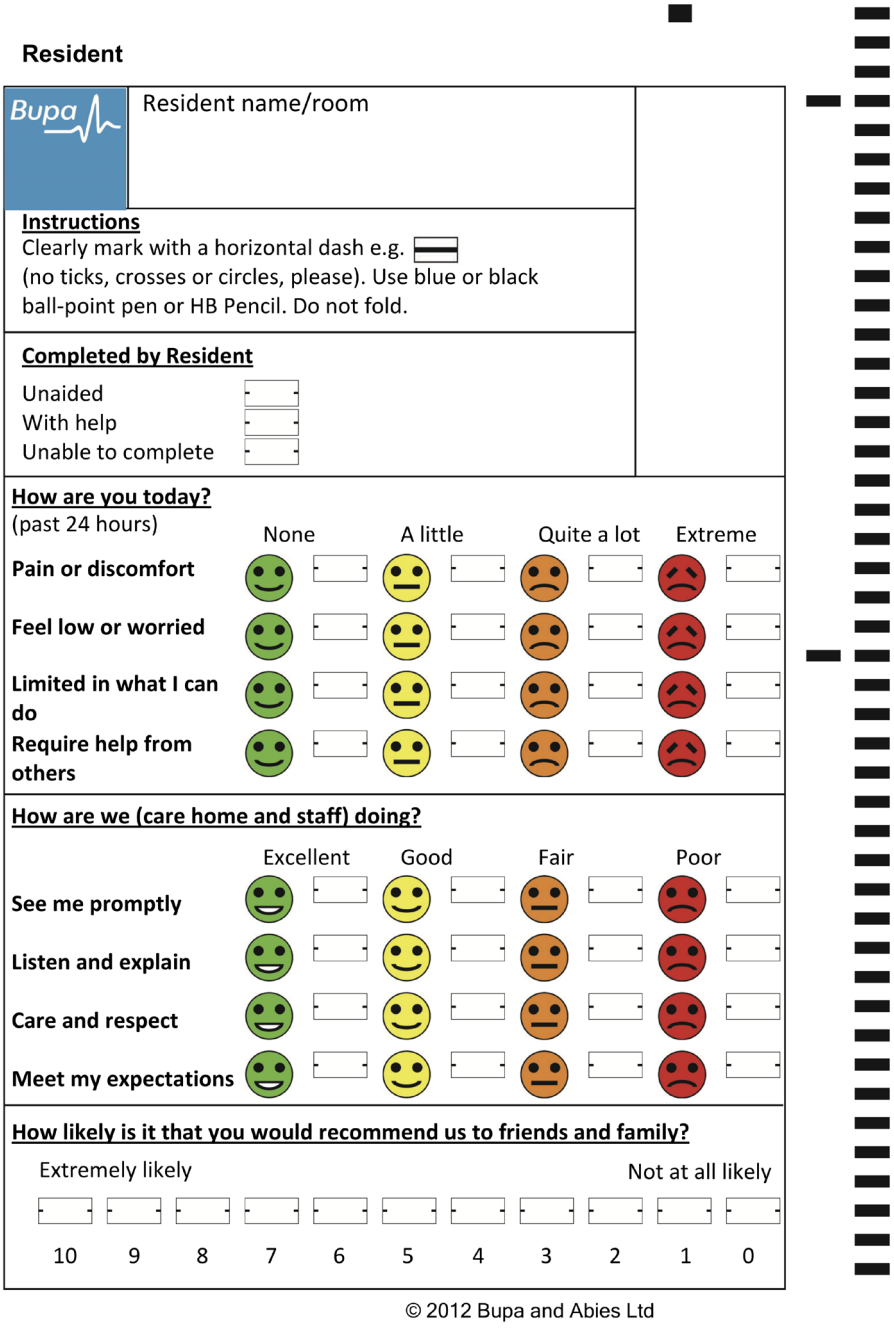

Figure 1 Resident form.

answered by staff is *How is the resident today?*. *HowRu* has four items:

- *Pain or discomfort*—physical symptoms.
- *Feel low or worried*—distress and emotional symptoms.
- *Limited in what I (he/she) can do*—disability, activities of daily living and leisure activities.
- *Require help from others*—dependency and self-care.

Each item has four possible responses: *Extreme*, *Quite a lot*, *A little* and *None*. At the individual level, these are scored from 0 (*Extreme*) to 3 (*None*). The summary *howRu* score is the sum of the item scores, giving a scale with 13 possible values with a range from 0 (4×*Extreme*) to 12 (4×*None*).

At the aggregate level, used here, all scores are transformed to a scale from 0 to 100. Individual item scores are multiplied by 100 and divided by 3; individual summary scores are multiplied by 100 and divided by 12. Using a common 0–100 scale aids understanding and comparison.

This analysis uses all the returns with complete paired ratings for all four *howRu* domains.

Responses for each region were collated regionally and forwarded to a central scanning bureau for data entry. Data for both staff and resident forms were entered centrally by scanning the optically marked forms. The data were imported into a database and exported to

Table 1 Overall distribution of staff (S) and resident (R) paired ratings of *howRu* items (n=10 168)

	Rater	None	A little	Quite a lot	Extreme
Pain or discomfort (*discomfort*)	S	5450 (54%)	3749 (37%)	879 (9%)	90 (1%)
	R	5403 (53%)	3529 (35%)	1122 (11%)	114 (1%)
Feeling low or worried (*distress*)	S	4998 (49%)	3868 (38%)	1131 (11%)	171 (2%)
	R	5249 (52%)	3606 (36%)	1106 (11%)	207 (2%)
Limited in what he/she can do (*disability*)	S	1352 (13%)	3599 (35%)	3578 (35%)	1639 (16%)
	R	1525 (15%)	3616 (34%)	3501 (36%)	1526 (15%)
Requires help from others (*dependence*)	S	814 (8%)	3325 (33%)	3804 (37%)	2225 (22%)
	R	1194 (12%)	3448 (34%)	3615 (36%)	1911 (19%)

Excel and the JASP statistical package (version 0.11) for analysis.[16]

We examined the overall distribution of results for staff and resident ratings. Differences in mean scores were tested using the Wilcoxon signed rank test. Correlations between staff and resident ratings were assessed using the Spearman rank correlation (r_s) and Cohen's kappa coefficient (κ). Kappa is a measure interobserver agreement that takes into account that raters will sometimes agree by chance.[17] No adjustments for multiple testing were made. The level of agreement was measured in terms of exact and ±one class.

Bias is the difference between two methods measuring the same things, such as a rating by staff (S) and self-rating by the resident (R). Here, bias is defined as resident score minus staff score (R–S).

Table 2 Distribution of *howRu* summary scores for staff proxy and resident self-report (n=10 168)

howRu score (0–12 scale)	howRu score (0–100 scale)	Staff proxy N (%)	Resident self-report N (%)
0	0	24 (0.2%)	18 (0.2%)
1	8.3	25 (0.2%)	45 (0.4%)
2	16.7	131 (1.3%)	128 (1.3%)
3	25.0	253 (2.5%)	257 (2.5%)
4	33.3	651 (6.4%)	578 (5.7%)
5	41.7	871 (8.6%)	828 (8.1%)
6	50.0	1577 (15.5%)	1464 (14.4%)
7	58.3	1377 (13.5%)	1311 (12.9%)
8	66.7	1783 (17.5%)	1757 (17.3%)
9	75.0	1265 (12.4%)	1275 (12.5%)
10	83.3	1253 (12.3%)	1350 (13.3%)
11	91.7	465 (4.6%)	545 (5.4%)
12	100	493 (4.8%)	612 (6.0%)

Bland and Altman, in a highly cited paper,[18] point out that reliance on mean scores and correlation does not mean that two methods agree. For example, mean scores may agree if bias is positive at high values and negative at low values; high correlations may be found when one measure is biassed consistently throughout its range or if bias is directly associated with value. They propose a method that plots bias (the difference between the two methods, (R – S)) against the average of the two methods (R+S)/2.

Our data differ from that envisaged by Bland and Altman. (1) We have a large number of paired measurements, which means that it is not feasible to plot individual points. (2) Our data are categorical ordinal data, with a limited number of categories (not interval or ratio continuous data), so that individual categories contain hundreds or thousands of instances. However, we find that mean scores, and hence mean bias, can be treated as if they are interval with few problems.

We plot the overall mean bias between the two methods (mean (R – S)) on the y-axis, against the actual average scores ((R+S)/2) for each instance on the x-axis. The number of categories on the x-axis is (2m–1), where m is the number of possible categories for each measure. For example, each item has four possible values, so the number of possible average scores is 7. The *howRu* summary score has 13 possible values (0–12 inclusive), so there are 25 possible average scores. The bias for the floor and ceiling average scores is always zero.

To summarise:

- Mean bias=mean (R – S).
- Actual average score = (R+S)/2.

In addition, we also show the number of responses for each actual average score. This distribution is not normal, because the number of paired ratings showing exact agreement for any item is larger than the numbers showing non-agreement.

Ethics statement

We carried out secondary analysis of data collected as part of a routine census of care home residents. The data

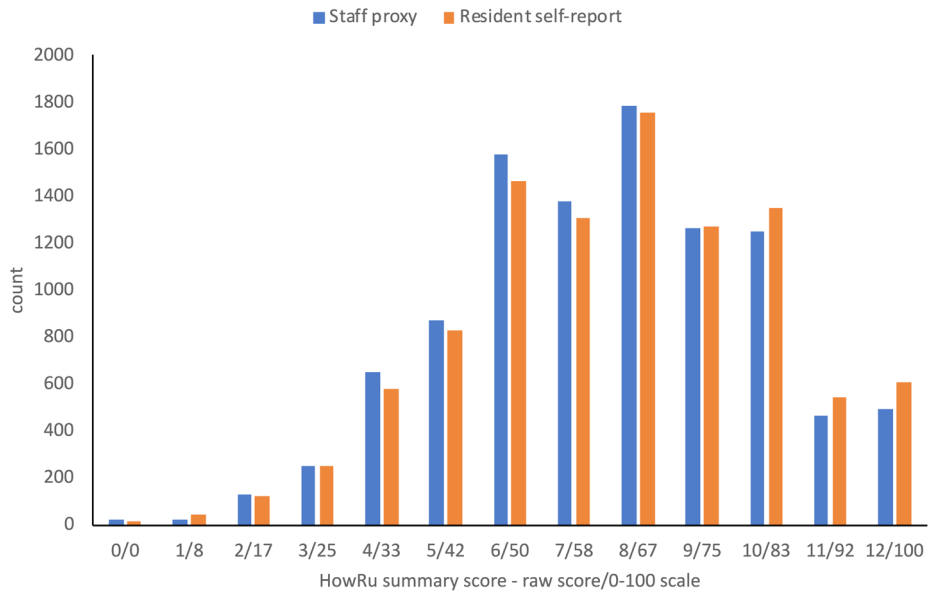

Figure 2 Distribution of *howRu* summary scores rated by staff proxies and residents.

were anonymous and undertaken to evaluate the current services without randomisation, so ethics approval was not required or sought. No data were collected about identifiable people and there was no risk to individual residents.[19]

Patient and public involvement

Care home residents and staff collected the data as part of a census to collect management information. All data were anonymous.

RESULTS

The census covered 24 506 residents in 395 care homes across UK, Australia and New Zealand. A total of 19 438 responses were received, of which 18 615 had health status data completed by staff and 10 712 by resident self-report. Ten thousand one hundred sixty eight responses (54.6% of staff ratings) included complete health status data rated by staff and resident self-report. This paper uses this data set.

Table 1 shows the distribution for each paired *howRu* item of staff proxy (S) and resident self-report (R) ratings. The distributions of *discomfort* and *distress* are broadly similar but differ considerably from those of *disability* and *dependence*, which are also broadly similar.

The distribution of *howRu* summary scores for staff and resident ratings is shown in table 2 and figure 2. Staff and residents generated the same summary score for 39.1% of residents and gave the same scores on all four items for 32.9%.

Table 3 shows the Spearman correlation, kappa and percentage of exact and plus or minus one class agreement between paired staff and resident self-ratings for each *howRu* item and the summary *howRu* score.

Spearman correlations for both items and the summary score are between r_s=0.54 and r_s=0.67, which may be interpreted as moderate or strong. For items, kappa is between κ=0.43 and κ=0.53, which may be interpreted as being moderate. For the summary score, κ=0.31, which may be interpreted as being fair. For items, the percentage of exact agreement is between 59.8% and 68.9% and agreement within one class is between 92.9% and 95.9%. For the summary score, exact agreement is 39.1% and agreement within one class is 66.0%. This is acceptable given 12 df.

However, distributions, correlations and exact agreement do not tell the whole story.

Table 3 Spearman's correlation, Cohen's kappa and levels of agreement between staff and resident ratings for *howRu* items and summary score (n=10 168)

Item	df	Spearman's correlation	Cohen's kappa	% exact agreement	%±1 class agreement
Pain or discomfort (discomfort)	3	0.55	0.46	68.9%	95.9%
Feeling low or worried (distress)	3	0.54	0.51	66.8%	95.3%
Limited in what you can do (disability)	3	0.61	0.43	59.8%	92.9%
Require help from others (dependence)	3	0.67	0.53	64.4%	94.5%
HowRu summary score	12	0.66	0.31	39.1%	66.0%

Table 4 Mean scores, SD, SE of the mean (SEM) and confidence limits (CL) for staff (S), resident (R), mean bias (R–S) and mean score (R+S)/2 for *howRu* items and summary score (n=10 168)

Item		Mean	SD	SEM	95% CL
Pain or discomfort	Staff proxy (S)	81.1	22.9	0.23	80.6 to 81.5
(*discomfort*)	Resident (R)	80.0	24.2	0.25	79.5 to 80.4
	Mean bias (R–S)	–1.11*	22.4	0.30	–1.5 to –0.7
	Mean score (R+S)/2	80.5	20.7	0.29	80.1 to 80.9
Feeling low or worried	Staff proxy (S)	78.2	24.7	0.19	77.7 to 78.7
(*distress*)	Resident (R)	78.9	25.2	0.24	78.4 to 79.4
	Mean bias (R–S)	0.67	23.5	0.25	0.2 to 1.1
	Mean score (R+S)/2	78.6	22.0	0.30	78.1 to 79.0
Limited in what you can do	Staff proxy (S)	48.6	30.5	0.30	48.0 to 49.2
(*disability*)	Resident (R)	50.2	30.7	0.20	49.6 to 50.8
	Mean bias (R–S)	1.56*	27.3	0.22	1.0 to 2.1
	Mean score (R+S)/2	49.4	27.4	0.23	48.9 to 49.9
Require help from others	Staff proxy (S)	42.3	29.7	0.27	41.7 to 42.9
(*dependence*)	Resident (R)	46.2	30.7	0.25	45.6 to 46.8
	Mean bias (R–S)	3.92*	24.8	0.16	3.4 to 4.4
	Mean score (R+S)/2	44.2	27.5	0.21	43.7 to 44.8
Health status summary score	Staff proxy (S)	62.5	19.5	0.22	62.2 to 62.9
(*howRu*)	Resident (R)	63.8	20.0	0.27	63.4 to 64.2
	Mean bias (R–S)	1.26*	16.4	0.27	0.9 to 1.6
	Mean score (R+S)/2	63.2	18.0	0.18	62.8 to 63.5

*Statistically different, Wilcoxson signed-rank test p<0.001.

Table 4 shows for each *howRu* item and the summary score the mean, SD, SE of the mean (SEM) and 95% confidence limits. We show staff proxy ratings (S), resident self-ratings (R), mean bias (R–S) and mean of staff and resident scores ((R+S)/2).

Figure 3 shows the mean bias (R–S) for each mean score for each of the four items (left hand axis), together with the percentage of responses for each mean score value (right-hand axis).

Residents score *Pain and discomfort* worse than staff when it is bad, but not when they have little or no pain or discomfort. The average bias (R–S) is –1.11. Residents rate *Feeling low or worried* somewhat worse than staff do when it is bad but somewhat better than staff when happier. Overall, the

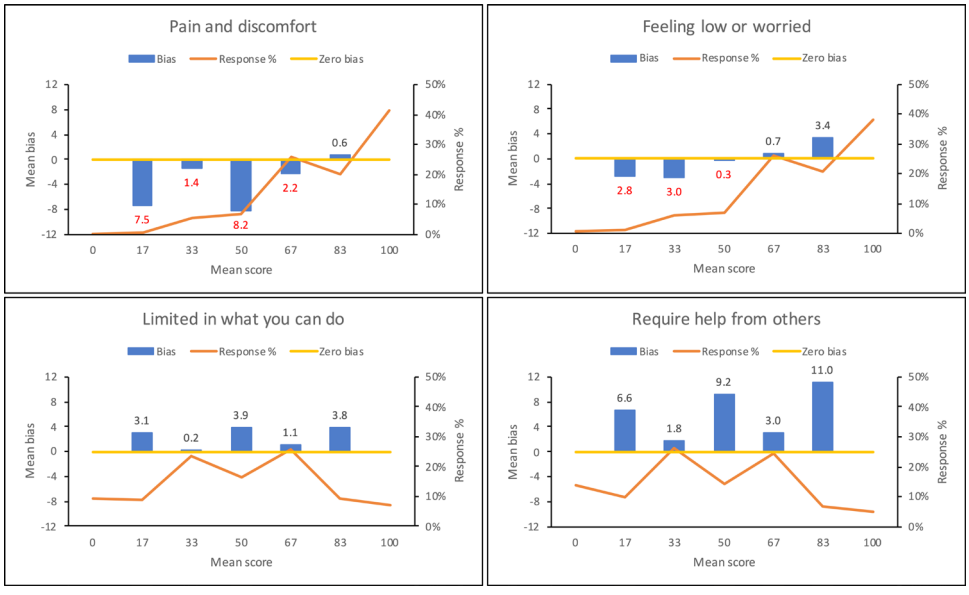

Figure 3 Bias and distribution of ratings for the *howRu* summary score (n=10 168).

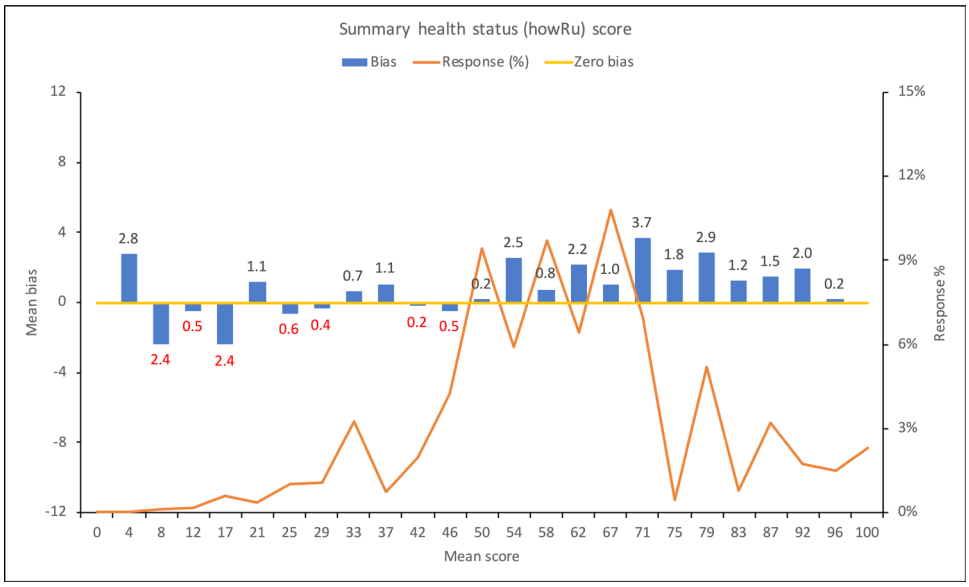

Figure 4 Bias and distribution of ratings for the *howRu* summary score (n=10 168).

average bias (R–S) is 0.67 (not significant, Wilcoxson signed-rank test p=0.054). Residents rate *Limited in what you can do* as somewhat higher (better) than do staff. The average bias (R–S) is 1.56. Residents rate *Require help from others* as substantially higher (better) than do staff. The average bias (R–S) is 3.92. The health status summary score is higher for residents than for staff. The average bias (R–S) is 1.26.

Figure 4 shows the mean bias (R–S) for the *howRu* summary score (left-hand axis), together with the percentage of responses for each mean score value (right-hand axis). Residents tend to rate themselves as having somewhat better health status than do staff, although the picture varies across the range. At the lower (worse) end, residents tend to score themselves lower than staff, while at the higher (better) end, residents score themselves as better than staff do.

DISCUSSION

This is the largest study (n=10 168) of matched ratings by care home staff and residents (or patients) that we are aware of. The size of the data set means that our estimates for mean scores for this population are quite precise.

Correlations are moderate or high and levels of exact agreement are satisfactory. We found differences in the distribution of bias for each item and the overall summary score.

The distribution of health status ratings by staff differs from resident self-rating overall; these differences also differ for each dimension of health status. Bland and Altman's contention, that differences in mean scores correlations and exact agreement rates can miss important aspects of bias such as an association with value,[18] is shown to be valid in the case of care home residents.

The probable explanations differ for each item.

Assessing how another person is feeling in terms of pain and distress is difficult. Residents may appear free of pain and distress, for example, when engaged in an activity yet suffer badly from night cramps or simply be low in mood or feel unhappy about loss of independence. Care home staff build their assessments from direct interaction and more general observation as well as from more formal assessment questioning. While we cannot determine which perspective should prevail, it may be that systematic robust staff observations using a PROM could avert unnecessary medication and consequent side effects.

Care home staff have broad day-to-day experience and judge disability and dependence in the context of people outside the care home. This may be more realistic, in a broader context, than the views of residents, who may reference their disability and dependence against that of other residents. This may lead them to believe that they can do more for themselves than they really can. Other residents may have little awareness of their limitations as a consequence of cognitive impairment.

The minimally important difference (MID) provides a measure of the smallest change that people regard as important.[20] Ideally, an anchor-based approach is most appropriate, but in the absence of a suitable anchor, MID can be estimated using a distribution-based method. At the individual level, half a SD is a widely used criterion.[21] For populations, the 95% CI = ±1.96(SD/√n). Sample size (n) is a critical factor. In practical terms, these tools are likely to be used to monitor the performance of care homes, or units within larger homes. For example, if a care home has 25 residents and SD=25 (see table 4), then the 95% CI is approximately ±10 and an appropriate MID threshold is ±5.

Red amber green (RAG) rating is widely used in quality control and improvement work. It could be used with *howRu* as follows. If a care home is monitoring staff-reported *howRu* scores on a weekly or monthly basis, a change of less than five points in the mean score would be rated green. Between 5 and 10 points would be rated amber and should be reviewed. More than 10 points should be rated red and trigger immediate investigation to understand what is going on.

This study has used secondary analysis of data to examine the relationship between staff and resident self-report ratings of health status in care homes. The study was not originally conceived for this purpose. This analysis excludes residents who staff considered could not or should not self-complete the ratings, such as people with advanced dementia or close to end of life. A possible risk was that staff encouraged residents to give the same answers as themselves for all four items. The central team had little control over this. This took place for 32.9% of residents, which does not seem too high.

The distributions of staff and resident ratings are similar superficially but differ in detail. Correlations between matched ratings for item and summary scores are moderate or strong. For items, more than 92.9% of paired responses are within plus or minus one class; for the *howRu* summary score, 66% are within plus or minus one class. Mean bias (resident minus staff scores) on 0–100 scale are discomfort (−1.11), distress (0.67), discomfort (1.56), dependence (3.92) and for summary *howRu* score (1.26).

CONCLUSIONS

We have demonstrated the differences between resident and staff assessments of their health status at scale using *howRu*.

Residents rated discomfort and distress lower (worse) than staff at severe levels, with bias associated with value. Residents rated their own disability and dependence as higher (better) than did staff, with bias not associated with value.

Staff may be better able to assess care home resident health status than can most residents, but may need training and take care not to underestimate severe pain and distress.

Tracking individual resident scores may provide a means to support residents proactively. RAG thresholds may provide a simple method to monitor changes in care home performance from a resident perspective.

Acknowledgements The authors are grateful for the support of Bupa Care Services, which funded the census, and to the care home staff and managers who took part. They are grateful to Henry Potts of the UCL Institute of Health informatics for statistical advice and for introducing them to the Bland-Altman method.

Contributors TB designed the surveys with CB and wrote the first draft of the paper. TB performed the analyses. Both authors managed the data collection, contributed to the final text, read and approved the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests TB is a director and shareholder of R-Outcomes, which provides quality improvement and evaluation services in the health and social care sectors using *howRu*. CB was previously medical director of Bupa Care Services and is a non-executive director of AKARI Care Homes, FINCCH and Invatech Health, all of which have interests in care homes and social care.

Patient consent for publication Not required.

Ethics approval The authors carried out secondary analysis of data collected as part of a routine census of care home residents. The data were anonymous and undertaken to evaluate the current services without randomisation, so ethics approval was not required or sought. No data were collected about identifiable people and there was no risk to individual residents.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Tim Benson http://orcid.org/0000-0002-2101-1353

REFERENCES

1 Prince M, Knapp M, Guerchet M, *et al*. *Dementia UK*. 2nd edn. London: Alzheimer's Society, 2014.
2 Crespo M, Bernaldo de Quirós M, Gómez MM, *et al*. Quality of life of nursing home residents with dementia: a comparison of perspectives of residents, family, and staff. *Gerontologist* 2012;52:56–65.
3 Moyle W, Murfield JE, Griffiths SG, *et al*. Assessing quality of life of older people with dementia: a comparison of quantitative self-report and proxy accounts. *J Adv Nurs* 2012;68:2237–46.
4 Devine A, Taylor SJC, Spencer A, *et al*. The agreement between proxy and self-completed EQ-5D for care home residents was better for index scores than individual domains. *J Clin Epidemiol* 2014;67:1035–43.
5 Clare L, Quinn C, Hoare Z, *et al*. Care staff and family member perspectives on quality of life in people with very severe dementia in long-term care: a cross-sectional study. *Health Qual Life Outcomes* 2014;12:175.
6 Usman A, Lewis S, Hinsliff-Smith K, *et al*. Measuring health-related quality of life of care home residents: comparison of self-report with staff proxy responses. *Age Ageing* 2019;48:407–13.
7 Centre for policy on ageing. A profile of residents in BUPA care homes: results from the 2012 BUPA census, 2012. Available: http://www.cpa.org.uk/information/reviews/Bupa-Census-2012.pdf [Accessed 6 Aug 2019].
8 Benson T, Sizmur S, Whatling J, *et al*. Evaluation of a new short generic measure of health status: *howRu*. *Inform Prim Care* 2010;18:89–101.
9 Benson T, Bowman C. Health status of care home residents: practicality and construct validity of data collection by staff at scale. *BMJ Open Qual* 2019;8:e000704.
10 Benson T. Measure what we want: a taxonomy of short generic person-reported outcome and experience measures (PROMs and PREMs). *BMJ Open Qual* 2020;9:e000789.
11 Hendriks SH, Rutgers J, van Dijk PR, *et al*. Validation of the howRu and howRwe questionnaires at the individual patient level. *BMC Health Serv Res* 2015;15:447.
12 Benson T, Potts HWW, Whatling JM, *et al*. Comparison of howRU and EQ-5D measures of health-related quality of life in an outpatient clinic. *Inform Prim Care* 2013;21:12–17.
13 Benson T, Williams DH, Potts HWW. Performance of EQ-5D, howRu and Oxford hip & knee scores in assessing the outcome of hip and knee replacements. *BMC Health Serv Res* 2016;16:512.
14 Benson T, Potts HWW. A short generic patient experience questionnaire: howRwe development and validation. *BMC Health Serv Res* 2014;14:499.
15 Reichheld F. *The Ultimate Question*. Harvard Business School Press: Boston MA, 2006.
16 JASP Team. JASP (Version 0.11) 2019 [Computer software].
17 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
18 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
19 NHS Health Research Authority. *Defining research: research ethics service guidance to help you decide if your project requires review by a research ethics Committee*. UK Health Departments' Research Ethics Service;, 2016.
20 Johnston BC, Ebrahim S, Carrasco-Labra A, *et al*. Minimally important difference estimates and methods: a protocol. *BMJ Open* 2015;5:e007953.
21 Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582–92.