

Health of the Nation Outcome Scales (HoNOS)

Research and development

J. K. WING, A. S. BEEVOR, R. H. CURTIS, S. B. G. PARK, S. HADDEN and A. BURNS

Background An instrument was required to quantify and thus potentially measure progress towards a Health of the Nation target, set by the Department of Health, "to improve significantly the health and social functioning of mentally ill people".

Method A first draft was created in consultation with experts and on the basis of literature review. This version was improved during four stages of testing: two preliminary stages, a large field trial involving 2706 patients (rated by 492 clinicians) and tests of the final Health of the Nation Outcome Scales (HoNOS), which included an independent study ($n=197$) of reliability and relationship to other instruments.

Results The resulting 12-item instrument is simple to use, covers clinical problems and social functioning with reasonable adequacy, has been generally acceptable to clinicians who have used it, is sensitive to change or the lack of it, showed good reliability in independent trials and compared reasonably well with equivalent items in the Brief Psychiatric Rating Scales and Role Functioning Scales.

Conclusions The key test for HoNOS is that clinicians should want to use it for their own purposes. In general, it has passed that test. A further possibility, that HoNOS data collected routinely as part of a minimum data set, for example for the Care Programme Approach, could also be useful in anonymised and aggregated form for public health purposes, is therefore testable but has not yet been tested.

The Government White Paper *Health of the Nation* (Department of Health, 1992, 1993) identified five key areas, one of which was mental health, where priority should be given to the development of local strategies for reducing mortality and morbidity. One of the mental health targets was "to improve significantly the health and social functioning of mentally ill people". This could be expressed either in terms of clinical and social improvement or by maintenance of an optimal functional state by preventing, slowing and/or mitigating deterioration. In an optimal functional state people with disabilities are able to function, if they wish to do so, reasonably close to the peak of their intact abilities. This is also close to the concept of 'optimal functional autonomy'. In order to set a quantified target and to measure how far it was being met, a dedicated instrument would be needed that could be used routinely, within the context of other necessary information, in the National Health Service (NHS).

PLAN OF DEVELOPMENT

To be useful to clinicians, the instrument would need to be brief enough for routine use by keyworkers, cover common clinical problems and social functioning, be sensitive to change or lack of it and have known reliability and relationship to more established scales. If successful, Health of the Nation Outcome Scales (HoNOS) data could be used as part of a minimum data set containing information on diagnosis, treatment, settings and background population, within which the indicators could be properly interpreted, comparing like with like. The creation of this essential context was not, however, part of the remit of this study.

Reviews of the literature revealed a huge number of scales but none met the chief criterion, which was that they must be brief enough for routine use across the

country while also covering the clinical and social range required. A first mock-up was therefore created, with the help of consultants, and tested and modified during four phases of development (Table 1).

TWO PILOT PHASES

HoNOS-1 was a 20-item instrument, drafted with the help of consultants and covering four key areas of functioning: behaviour, impairment, symptoms and social functioning. Item 20 was a global (0–100) disability score. This version was tested during a first pilot study ($n=152$) to discover how it performed in terms of clinical acceptability, simple structure and sensitivity to change. The results were satisfactory. In the light of comments from users it was shortened to 12 items, each rated on a 0–4 scale of severity. HoNOS-2 was then tested again ($n=100$) with results closely similar to those of the first pilot. HoNOS-3 (see Table 2), used in the main field trials, reflected the lessons learned. A glossary based on this experience was provided containing a definition at each rating point on each of the 12 items. A set of answers to common questions raised by raters and a description of the basic rationale and methods of using the system was included.

FIELD TRIALS – HoNOS-3

Large-scale trials were needed in order thoroughly to test feasibility and clinical acceptability and establish a substantial base of data from ratings made by clinicians as part of their everyday work. The community teams taking part were recruited partly from those already involved in the first two stages, partly by contacts through the Royal College of Psychiatrists' Research Unit (CRU), which holds a list of district audit convenors, and partly from volunteers who had heard of the HoNOS projects. A choice was made on the basis of contrasting service and staffing patterns, geography and socio-demographic indices. District rankings on the Jarman index of social deprivation (Jarman & Hirsch, 1992), for example, ranged from 8 to 188 out of the total 192 districts.

The task in each area was routinely to collect HoNOS data on a consecutive series of up to 200 patients (T1), and to follow as many as possible for at least three months or to the end of the episode of contact, if

Table 1 HoNOS project: phases of development

Phase of work (Dates)	1. Start-up (Sep. 92–Mar. 93)	2. Pilot tests (Apr. 93–Dec. 93)		3. Field trials (Jan. 94–Mar. 95)	4. Reliability/comparisons (Apr. 95–Sep. 95)
Version of HoNOS	–	HoNOS-1	HoNOS-2	HoNOS-3	Final HoNOS
Sites (n)	–	9	7	25	2 (main) 4 (supplementary)
Patients (n)	–	152	100	2706	Reliability 293+188 Comparisons 166+107
Nature of work	Literature search Consultation Design of trials	Acceptability Structure Sensitivity		Acceptability Structure Sensitivity Reliability Profiles	Acceptability Structure Sensitivity Reliability Comparisons Profiles

earlier (T2). Local supervisors and coordinators were trained on-site. Completed score sheets were returned to the CRU for checking and recording before being forwarded in batches for data entry.

The statistical analyses were based on data from the charts and score sheets of 2706 patients from 25 sites. Of these, 1678 were rated on a second occasion (872 by the same rater on both occasions and 806 by a different rater on the second occasion), making 4380 sets of data in all. The same issues (clinical acceptability, simple structure and sensitivity to change) were studied as in the pilots but the larger numbers allowed more complexity in analysis. Checks were made on the proportions of missing item values. For example, of the 872 same-rater pairs, 6.4% contained missing data in at least one item. There was little effect on the histograms of distributions or on the means. The analysis for different-rater pairs gave similar results.

Analysis of HoNOS-3 structure

It was intended that, in a 12-item instrument covering a broad socio-clinical area, each item should be informative. High internal consistency was not sought. The 12 × 12 item matrix was satisfactory in this respect in that it showed only three correlations with values of 0.30 or higher. The four sub-scores (A–D) were no more intercorrelated (0.08–0.25) than the individual items. The correlations between A–D sub-scores and the total score were 0.62, 0.45, 0.63 and 0.72. The equivalents for the global score were 0.06, 0.37, 0.33 and 0.28. Principal components analysis of

data at T1 with varimax rotation gave five factors accounting for 63% of the variance. These were fairly close to the grouping into sections A–D adopted on clinical grounds.

The relationship between ratings of severity on individual items and the total score (ratings on items 1–11 summed) is important if a single indicator of severity is to be derived. Table 3 shows data derived

from HoNOS-3 and also provides results from a subsequent check using the final HoNOS. The index is based on the fact that each item rating represents a clinical judgement of severity: the more zeros there are, the less the severity; the more fours the greater the severity; and *pro rata* in between. The threshold for the highest level in the hierarchy shown in Table 3 is defined by four or more ratings of four (very severe),

Table 2 HoNOS-3 and the final HoNOS compared (all items scored 0–4)

HoNOS-3 items	Final HoNOS items
Sub-score A: max. 12	Sub-score A: max. 12
1. Aggression	1. Aggression & overactivity
2. Self-harm	2. Self-harm
3. Substance use	3. Substance use
Sub-score B: max. 8	Sub-score B: max. 8
4. Cognition	4. Cognition
5. Physical health	5. Physical health
Sub-score C: max. 12	Sub-score C: max. 12
6. Affective disorders	6. Hallucinations & delusions
7. Psychotic disorders	7. Depression
8. Other symptoms	8. Other symptoms
Sub-score D: max. 12	Sub-score D: max. 16
9. Social relations	9. Social relations
10. Housing	10. General functioning
11. Activities	11. Housing
12. —	12. Activities
Total score: max. 44	Total score 1–12: max. 48
Global score (0–100)	Total score 1–10: max. 40 ¹

1. If relevant social information is not available, items 11 and 12 should be omitted from the total score and Total 1–10 used instead.

Table 3 Item severity ratings at Time 1 by total scores: HoNOS-3 and final HoNOS

Minimal value	HoNOS-3		Final HoNOS	
	Mean=9.86	n=2612	Mean=9.98	n=641
4 or more 4s	24.2	41	22.5	6
3 × 4s	19.5	44	20.4	16
2 × 4s	16.2	148	16.6	47
1 × 4	12.4	487	13.8	99
4 or more 3s	17.9	57	17.7	27
3 × 3s	14.7	103	13.7	24
2 × 3s	11.8	263	10.9	71
1 × 3s	8.8	573	8.1	134
4 or more 2s	10.8	66	10.3	25
3 × 2s	8.4	112	7.9	38
2 × 2s	6.4	205	5.3	55
1 × 2	4.7	264	3.5	67
4 or more 1s	4.9	54	4.7	3
3 × 1s	3.0	51	3.0	6
2 × 1s	2.0	63	2.0	6
1 × 1	1.0	44	1.0	9
only 0s	0.0	37	0.0	9

that is, a score of 16. In fact, other items are rated as well, giving a mean of 24.2. Because of this, there is an acceptable overlap in some of the means shown. Thus, four or more ratings at level three suggest higher total severity than two ratings at level four,

three ratings at level three suggest more severity than one at level four, and so on.

Sensitivity to change

The main field trials provide evidence of change between first and second ratings (T1

and T2) and how far this is supported by the clinical judgement of the rater at T2. Change in individual item scores is indicated by the profiles of means at T1 and T2, shown in Figure 1 for the same and different pairs of raters. Items six, eight and nine have the highest mean ratings on both occasions. The overall pattern of improvement on the second occasion of rating is evident.

Table 4 shows the means for sub-scores A–D, the total score derived from items 1–11 (0–44) and the global item 12 (0–100), at T1 and T2. All means show a substantial and highly significant improvement, whether derived from same- or different-rater pairs. The largest mean change per item is found in the symptom sub-score, followed by behaviour, social functioning, and impairment, in that order. The sub-scores thus provide a degree of prediction about likely change.

Table 4 also shows, what is clear from inspection of Figure 1, that the means from different-rater pairs were all higher at T1 than those from same-rater pairs. A possible reason for the interactions is that, because of the shift system, different-rater pairs constitute 65% of those in acute wards compared with 12% in non-ward settings. This is illustrated in Table 5, which also shows that the higher severity in different-rater pairs at T1 occurs, though to a lesser extent, in non-acute settings as well. A confirmatory factor analysis, carried out separately for short-stay in-patients and others, showed that the underlying latent variable for same-rater pairs remained consistent over the two occasions in both kinds of setting. Changes in sub-scores between T1 and T2 correlated well with changes in total score, whether calculated using same-rater or different-rater pairs. Correlations with the global score were generally low.

To provide a check on the changes in score over time, raters on the second occasion were asked to make a clinical estimate of change during the period between ratings using a scale of ‘much better’, ‘better’, ‘no change’, ‘worse’ and ‘much worse’. Table 6 shows the mean change between total HoNOS scores at T1 and T2 for each of these intervals. A one-way analysis of variance for each type of pair showed a very highly significant relationship for both between the overall outcome estimated by retrospective clinical judgement and the change in total score. Apart from the interval between ‘much better’ and ‘better’ for same-rater pairs, all

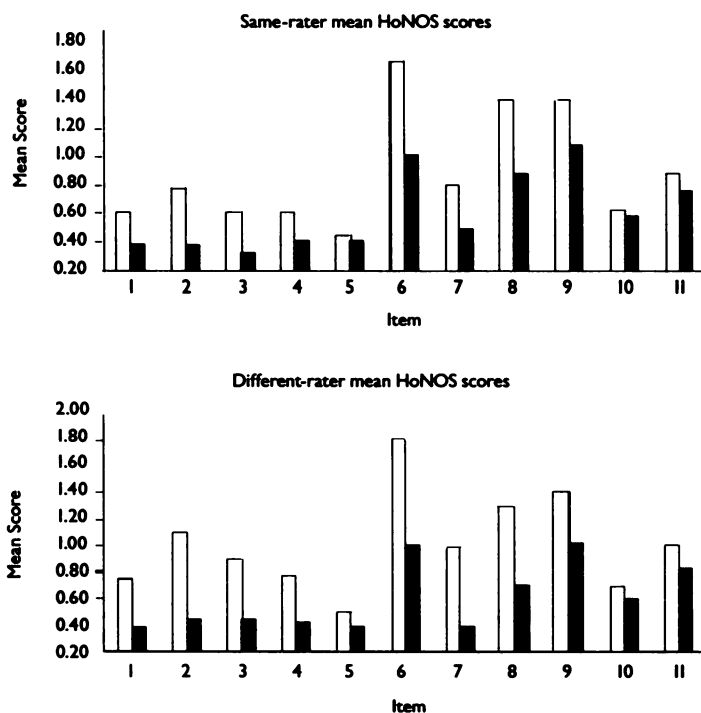


Fig. 1 Mean scores on items 1–11 at T1 and T2, by type of rater-pair. □, time 1 (T1); ■, time 2 (T2).

Table 4 HoNOS-3 sub-scores and summary scores at Time 1 and Time 2: by type of rater pairs

Type of score	Same raters ¹ (n=871)		Different raters ¹ (n=798)	
	T1	T2	T1	T2
A (0–12) Behaviour	1.97	1.08	2.62	1.35
B (0–8) Impairment	1.03	0.83	1.22	0.85
C (0–12) Symptoms	3.92	2.41	4.13	2.23
D (0–12) Social	2.91	2.43	3.04	2.45
Total (0–44)	9.77	6.67	10.96	6.86
Global (0–100)	43.37	33.65	44.37	34.10

1. Numbers vary because of missing data.

All T1–T2 differences are significant ($P < 0.001$). Difference between same-rater pairs and different-rater pairs: A $P < 0.001$; total $P < 0.001$; Interaction: A $P < 0.001$; B $P < 0.013$; C $P < 0.002$; total $P < 0.001$.

Table 5 Mean total scores at Time 1 and Time 2: by setting and type of rater pair

Time when rated	In-patient less than 3 months		All other patients	
	SRP (n=329)	DRP (n=611)	SRP (n=295)	DRP (n=39)
Time 1	10.93	11.05	8.59	10.28
Time 2	6.88	5.54	6.15	7.49
Improvement T1–T2	4.05	5.51	2.44	2.79

SRP, same-rater pairs; DRP, different-rater pairs.

Table 6 Changes in total score T1–T2, by retrospective judgement

Estimated outcome	Same-rater pairs		Different-rater pairs	
	Mean change	n	Mean change	n
Much better	5.69	156	6.13	193
Better	4.48	327	4.37	331
No change	0.66	194	2.22	155
Worse or much worse	–1.11	45	–2.2	25
Total	3.37	722	4.15	704

mean score changes between the raters' categories are individually significant. Although the second rater was not always blind to the earlier HoNOS ratings, this does lend extra clinical support (particularly in the case of different-rater pairs) for the suggestion that the scales measure change.

Test–retest reliability

Tests of reliability were deferred to the fourth phase of the project since it was necessary to ensure that the information was available for the final version of HoNOS. However, an opportunity to consider test–retest reliability is provided by a comparison of the 212 pairs of ratings

made by the same rater on two occasions for which the clinical estimate at T2 was zero (no change). This avoids the usual problems arising from marked changes in patients' item-profiles over time or the rater at T2 being different from the rater at T1. The mean interval for this group was 35.3 days and the gender and diagnostic spread was much the same as for the total series. The intraclass correlation coefficients for items, the total score (sum of items 1–11) and the global score were all between 0.74 and 0.88, except for aggression (0.61).

Rating styles

Nearly all raters were nurses ($n=399$) or doctors ($n=60$) and the statistical analysis is

heavily weighted by their data. There were 19 social workers and seven each of psychologists and occupational therapists. It is important to know whether, as separate professions, they have different rating styles and, if so, what the consequences are for the measurement of outcomes. Table 7 shows the mean total HoNOS score for short-stay in-patients and for those in other settings (generally in longer-term contact and the majority not in-patients), according to the profession of the rater.

The difference in mean total score between nurses and doctors rating short-stay in-patients is significant but relatively small (1.8) for acute wards and not significant for longer-term settings. Selective factors, such as duties connected with the Mental Health Act, could be part or all of the explanation. This may also explain part of the high mean total by the few social workers in the acute ward setting. But the comparatively high mean score of social workers in other settings (mostly day centres) is more remarkable. The effect is seen in all four sub-scores (i.e. both clinical and social items) and in both psychotic and affective disorders.

Clinical profiles

If profiles of HoNOS-3 ratings and/or scores did not illustrate associations that have already been clearly demonstrated by other research, the instrument would lose clinical validity. For example, items with the highest mean score should differ predictably between diagnostic groups. Thus, among the ICD–10 disorders, F0 (mainly dementias) should score highest on cognitive problems and physical disability; F1 (substance use) should be associated with item three; F2 (affective disorders) with depression, and so on. This holds true. Age category is found to be associated with different items in predictable ways. There is a gradient of aggression, highest under age 25, lowest over 65. Cognition, as expected, has the opposite pattern. Patients in hospital for less than three months have high mean scores on almost every item compared with people not in hospital care; the only exception is item eight, mostly representing anxiety and phobic conditions, the mean for which is slightly higher in those who are not in-patients. Items on aggression and substance use are higher in males than in females. Pre-onset social and role performance shows a clear gradient, from above average to markedly below

Table 7 Mean total scores by profession and setting

Staff category	In-patient < 3 months		Other settings		All settings	
	Mean	n patients	Mean	n patients	Mean	n patients
Nurse	10.46	1094	8.28	696	9.61	1790
Doctor	12.26	360	7.94	345	10.15	705
Social worker	17.30	10	13.81	73	14.43	83
Occupational therapist			11.56	25	11.56	25
Clinical psychologist			7.28	40	7.28	40

average, in association with five items: aggression, cognition, psychotic symptoms, housing and activities.

Amendments to HoNOS-3 resulting from the phase three field trials

Comments made during follow-up visits to 14 sites, responses to a request in the regular HoNOS newsletter, and returns of a questionnaire sent to all raters, yielded few suggestions for further change and most amendments were small. One concerned the conjunction of manic (rare) and depressive (common) symptoms in one item; a practice that was in any case against the spirit of the rule that the scales deal with symptoms and dysfunctions, not diagnoses. Overactivity was therefore included at item one, while manic and depressive delusions, hallucinations and bizarre behaviour were included with other psychotic symptoms placed at item six, with depression following as item seven. The global rating and its change score were not closely associated with items, sub-scores or total score. Raters found it difficult to switch from the 0–4 scales to the 0–100. They tended to avoid the lowest and highest sectors and to cluster ratings round the remaining anchor points, resulting in a more normal distribution. Since it was not entirely clear how the global scale was being used, it was dropped and a more specific definition of disability provided, in terms of activities of daily living rated on the usual 0–4 scale, with glossary definitions for each severity point. This was placed at item 10, followed by the social environment item split into two for residential and daytime environments, with clarified glossary definitions in terms of autonomy. These changes were incorporated in the final version of HoNOS (see

Table 2). Although the 0–48 score range appears to give less scope for measuring change than one of 0–100, the top and bottom sections of the global score were hardly used and the quality of the information conveyed is unknown. If a global scale is required, it can be used independently.

RELIABILITY AND RELATIONSHIP TO OTHER SCALES

Studies in Nottingham and Manchester

The fourth phase consisted of tests to show whether the final form of the instrument remained acceptable to clinicians, had comparable characteristics to its predecessor, performed well when subjected to independent tests of reliability, and had reasonable compatibility with larger instruments with a long track record. The studies were carried out in academic departments in Nottingham and Manchester, in consultation with the CRU team.

Design of the reliability trials

In each city, a preliminary pilot was carried out to assess training and conditions of testing, followed by a main reliability trial. The Biometrics Department at the Institute of Psychiatry advised that, for most reasonable target coefficient requirements, 200 pairs would be appropriate. Altogether, 293 patients were rated during the two phases of testing, 96 during the pilots, 197 during the main study. The main difference from the group of patients in the field trials was that more patients in Manchester were elderly, many suffering from dementia.

In Nottingham, the trainer and central rater was a psychiatrist, in Manchester an experienced mental health nurse. This

arrangement was adopted in order to study the performance of HoNOS under a broad range of training and working conditions and to represent the expertise of the professions principally involved.

Reliability

In a first rapid assessment in order to discover any major problems, nine patients were rated by the Nottingham trainer and 13 other clinicians, eight of whom were the patients' keyworkers. Comparing the score sheets of each rater against those of each of the other 12 provided 78 sets of ratings for one patient, resulting in 702 sets of ratings (351 rater-pairs) for all nine. A few nurses had to leave early and, with some missing data, the actual number of rater-pairs analysed is 325. The exercise was intended to be used to provide initial feedback on the way the 12 items were being rated. In fact, only three were unsatisfactory and were dealt with by clarification of the wording of the related anchor points in the glossary. Subsequent work proceeded on the basis of these revisions, which were also fed back to the Manchester trainer in time for the main study there.

In the Nottingham main trial, 100 patients were rated by the trainer, and by each patient's keyworker or the person most knowledgeable about the patient's clinical condition. The first data column of Table 8 shows the intraclass correlation coefficients (ICC) attained, the coefficients of which are good to very good,¹ apart from item 12 which is moderate. There was also good equivalence between results examined when numbers reached 60 and 100, indicating good structural stability. The ICC data for four sub-scores and two total scores are all very satisfactory.

In Manchester, having first secured the collaboration of local hospital and community teams, the nurse trainer attended routine ward rounds and team meetings in order to ensure the inclusion of a spread of disorders and settings. Nine consultants conducted interviews with their patients following a general introduction. Further discussion ensued after the patient had left. The nurse trainer observed, but did not take part in, these proceedings. HoNOS was then rated by the trainer and relevant consultant.

1. Agreement categories (Landis & Koch, 1977): 0.41–0.60, moderate; 0.61–0.80, good; 0.81+, very good.

Table 8 Intraclass correlation coefficients (ICC) for HoNOS-4. Tests of reliability: Nottingham and Manchester

HoNOS item	Main study in Nottingham	Main study in Manchester
	n=100	n=97
1. Aggression	0.97	0.80
2. Self-harm	0.88	0.92
3. Drug & alcohol use	0.99	0.61
4. Cognitive problems	0.81	0.92
5. Physical illness & disability	0.88	0.89
6. Hallucinations & delusions	0.87	0.92
7. Depression	0.84	0.89
8. Other symptoms	0.95	0.52
9. Relationships	0.74	0.78
10. Activities of daily living	0.71	0.90
11. Residential environment	0.83	0.72
12. Day-time activities	0.49	0.51
<i>HoNOS subscores</i>		
A: 1–3 Behaviour	0.89	0.74
B: 4–5 Impairment	0.87	0.95
C: 6–8 Symptoms	0.88	0.81
D: 9–12 Social	0.82	0.68
<i>HoNOS total scores</i>		
E: 1–12	0.86	0.77
F: 1–10	0.86	0.86

As in Nottingham, data collection was divided into a pilot and a main phase. In the pilot, ratings were provided from 87 trainer–doctor pairs. After minor adjustments on the basis of these results and suggestions from the Nottingham pilot, the trainer in Manchester rated a further 97 patients, each also rated by the relevant consultant. The second data column of Table 8 shows that the ICC coefficients are good to very good for all items but two (items 8 and 12), which are moderately good and acceptable.

Comparisons

An update of the earlier search, and reference to a similar survey by Andrews & Morris-Yates (1994), found no instrument of reasonable length that covered both the clinical and the social items in HoNOS. Two sets of scales, the 24-item Brief Psychiatric Rating Scale (BPRS-24) (Overall & Gorham, 1962; Ventura *et al*, 1993) and the Role Functioning Scale (RFS; Goodman *et al*, 1993), were selected because of a long track record and because, between them, they cover most of the content of the final HoNOS. However,

both instruments have a much wider range and the BPRS is intended to be completed at an interview with the patient, which is only an option for HoNOS. See Wing *et al* (1996) for papers describing the properties of each instrument.

In Nottingham, the trainer completed the comparison instruments for 33 of the patients in the reliability study. The product–moment correlation between the total HoNOS and total RFS scores was 0.65. That between the four social HoNOS items (9–12) and RFS 1–4 was 0.75. The correlation between total HoNOS and total BPRS scores was 0.84 and that between the clinical items HoNOS 1–8 and BPRS 1–15 (omitting the behavioural items 16–24) was 0.85.

The exercise in Manchester was more complicated and less appropriate for comparisons, since nine consultants rated a total of 97 patients. Five contributed too few BPRS and RFS forms for separate analysis; a total of 20. The other four between them provided a total of 77 score sheets. Of these, 37 (22 with a diagnosis of dementia) were contributed by one rater. The correlations between total HoNOS and total RFS scores, in order of magnitude, were 0.52, 0.64,

0.67 and 0.73. The equivalents for total HoNOS and total BPRS scores were 0.49, 0.57, 0.71 and 0.71.

Detailed information on individual items is provided for both samples in Table 9, which shows the symptomatic BPRS-24 items that are correlated at a level of 0.50 or higher with HoNOS items. Data from the four Manchester raters are combined. Bearing in mind the different diagnostic mix of the two patient groups, and the fact that there were four raters in Manchester, the results show reasonably good equivalence.

DISCUSSION

Meeting the remit

The first five necessary (though not sufficient) requirements for an instrument capable of routinely measuring changes in mental health outcomes have been met as well as they reasonably could be during the stages of construction. HoNOS is simple to use and generally clinically acceptable; it covers a broad range of clinical problems and social dysfunctions; it is sensitive to change or the lack of it over time; it has acceptable reliability and is compatible with longer and well-established instruments. It is now in the public domain and released for more widespread use and experiment.

Training users is simple. Most clinicians need and already collect and record such information, albeit not in standard format. Apart from the numbers participating in the trials (see Table 1), on average 15 clinicians at each of 40 sites have since been trained to use the final HoNOS and 55 have received the full trainers' course at the CRU. The scales and glossary are in the public domain. A software application (HoNOSoft) can be used to enter data rapidly, make simple analyses and print results or export them to statistical packages. Materials and manuals, and a full report on the research, can be obtained from the CRU.

Clinical use of HoNOS

The most obvious use for HoNOS is as a simple record of a patient's progress. It takes only a few minutes to complete a small 'peel-off' rating form, which can be stuck at once into the case file and updated as needed. If the instrument continues to be found useful in this way by individual clinicians, the quality of information

Table 9 Correlations between final HoNOS and BPRS-24 items, Nottingham and Manchester

HoNOS item	BPRS-24 item	Nottingham	Manchester
1. Aggression	B5 Hostility	0.87	0.60
	B10 Disorientation	0.52	–
2. Self-harm	B19 ¹ Suicidality	0.97	0.72
	B5 Hostility	–	0.60
	B3 Depression	0.59	–
	B4 Guilt	–	0.50
4. Cognition	B10 Disorientation	0.74	0.84
	B11 Conceptual disorganisation	0.61	–
	B14 Blunted affect	0.61	–
	B16 ¹ Mannerisms	0.55	–
	B18 ¹ Emotional withdrawal	0.51	–
6. Hallucinations & delusions	B7 Hallucinations	0.78	0.88
	B7 Unusual thought content	0.81	0.58
	B6 Suspiciousness	0.75	0.59
	B24 ¹ Distractability	0.66	–
	B21 ¹ Bizarreness	0.57	–
7. Depression	B3 Depression	0.91	0.87
	B4 Guilt	0.81	–
	B2 Anxiety	–	0.61
	B19 ¹ Suicidality	0.60	–
	B15 Tension	–	0.59
8. Other symptoms	B18 ¹ Emotional withdrawal	–	0.50
	B1 Somatic concern	0.54	–
	B15 Tension	0.51	–
	B2 Anxiety	0.49	–

NB: HoNOS items 3 (drug misuse), 5 (physical problems) and 9–12 (social functioning) are not represented in BPRS-24. 1. BPRS items 16–24 are behavioural.

recorded is likely to be acceptable and many other ways of using HoNOS data become feasible.

The key to these would be incorporation into a structured clinical data set, for example serving a clinical team's need to implement the Care Programme Approach. There would be opportunities to compare specific problems or interventions or settings, and to audit outcomes in the context of care plans, case mix and case load (Lelliot, 1994, 1995). Rationally derived episodes could be substituted for the current method based on artificial divisions between periods of in-patient and other forms of care (Glover, 1995a,b).

Possible administrative uses of HoNOS

Whether the total HoNOS score would prove suitable as a single indicator for administrative purposes (such as monitor-

ing the attainment of a quantified target) can only be assessed in a realistic context such as a sector or district register. The total has proved robust throughout the four stages of the trials and the four sub-scores may also be found useful. Profiles of items provide the most detailed information but much will depend on the nature of the comparisons to be made. The most important criterion, whatever the question being asked, is that like should be compared with like. The more detailed the database, the easier that will be.

The incorporation of clinical information, following aggregation and anonymisation, into sector or district registers used for administrative as well as clinical and epidemiological purposes, requires assurance not only of the confidentiality, quality and security of clinical data, but also of the integrity of its use (Andrews & Morris-Yates, 1994). A discussion of the issues raised for

HoNOS, and for computerised health records more generally, by comparisons that have been made involving the work of institutions and practitioners in various fields of professional work will be the subject of another paper.

CONCLUSION

HoNOS performs to specification for clinical purposes. It has an official entry in the NHS Data Dictionary and is recommended for further testing within the secondary mental health services. It is being tested in Australia, Denmark, Italy and Spain.

A version for children and adolescents has been completed, others for old age psychiatry and learning disability are under test, and one for mentally disordered offenders is being developed. If a comparable primary care HoNOS were devised, tested and widely used, the age-old problem of relating data collected from secondary health and social agencies to a relevant population base would be much simplified.

ACKNOWLEDGEMENTS

The research on which this report is based was supported by the Department of Health. This study could not have been mounted without the hard work and cooperation of site coordinators who organised start-up, training and feedback meetings, supervised the running of projects and data collection, disseminated information during the project and were in many cases HoNOS raters themselves. We are indebted to Professor Brian Everitt, Dr Graham Dunn and Mr Charlie Sharp for their advice and help with analysis and interpretation of data.

APPENDIX

HoNOS scales, HoNOSoft software and further details are available from the College Research Unit, 11 Grosvenor Crescent, London SW1X 7EE.

REFERENCES

- Andrews G. & Morris-Yates, A. (1994) Ethical implications of the electronic storage of medical records. In *Computers in Mental Health*, Vol. 1 (eds G. Andrews, T. B. Ustun, H. Dilling, et al), pp. 129–135. St Andrews: Churchill Livingstone/WHO.
- Department of Health (1992) *Health of the Nation*. White Paper. London: HMSO.
- (1993) *Key Area Handbook. Mental Health*. London: Department of Health.
- Glover, G. (1995a) The public health perspective. In *Measurement for Mental Health*, Chapter 6 (ed Wing, J. K.). London: College Research Unit.

— (1995b) Mental health informatics and the rhythm of community care. *British Medical Journal*, **311**, 1038–1039.

Goodman, S. H., Sewell, D. R., Cooley, E. L., et al (1993) Assessing levels of adaptive functioning: the Role Functioning Scale. *Community Mental Health Journal*, **29**, 119–131.

Jarman, B. & Hirsch, S. (1992) Statistical models to predict district psychiatric morbidity. In *Measuring Mental Health Needs* (eds G. Thornicroft, C. R. Brewin & J. K. Wing), pp. 62–80. London: Gaskell.

Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lalliot, P. (1994) Making clinical informatics work. Don't forget the doctor is the customer. *British Medical Journal*, **308**, 802–803.

— (1995) Mental health information systems. *Measurement for Mental Health. Contributions from the College Research Unit*, chapter 8, (ed. Wing, J. K.). London: College Research Unit.

Overall, J. E. & Gorham, D. R. (1962) The Brief Psychiatric Rating Scale. *Psychological Reports*, **10**, 799–812.

Ventura, J., Green, M. F., Shaner, A., et al (1993) Training and quality assurance with the BPRS. 'The drift busters'. *International Journal of Methods in Psychiatric Research*, **3**, 221–244.

Wing, J. K., Curtis, R. H. & Beevor, A. S. (1996) HoNOS. Health of the Nation Outcome Scales. Report on research and development. London: College Research Unit.

CLINICAL IMPLICATIONS

- HoNOS is suitable for routine use by nurses and psychiatrists because of its simplicity, broad clinical and social coverage and adequate psychometrics.
- Item profiles, subscores and totals can be used to measure progress.
- It would be suitable for incorporation into a data set for CPA monitoring.

LIMITATIONS

- The use of HoNOS by other keyworker groups has not been tested.
- The instrument is not intended to explain why the outcomes it measures have occurred.
- The research has provided an instrument that could be used within a broader data set to monitor progress towards local, regional and national targets. Further study is needed to test this proposition.

J. K. WING, FRCPsych, A. S. BEEVOR, R. H. CURTIS, College Research Unit, 11 Grosvenor Crescent, London SW1X 7EE; S. G. B. PARK, MRCPsych, University of Nottingham, Duncan Macmillan House, Porchester Road, Nottingham; J. HADDEN, A. BURNS, FRCPsych, University of South Manchester, Withington, Manchester M20 9BX

Correspondence: Professor J. K. Wing, College Research Unit, 11 Grosvenor Crescent, London SW1X 7EE

(First received 23 January 1997, final revision 12 August 1997, accepted 26 August 1997)
