Multilevel investigation of variation in HoNOS ratings by mental health professionals: a naturalistic study of consecutive referrals

R. ECOB,¹ T.J. CROUDACE,¹ I. R. WHITE,² J. E. EVANS,³ G. L. HARRISON,³ D. SHARP,⁴ P. B. JONES¹

¹ Department of Psychiatry, University of Cambridge

² MRC Biostatistics Unit, Cambridge

³ Department of Psychiatry, University of Bristol

⁴ Department of Social Medicine, University of Bristol

Abstract

Episodes of mental healthcare in specialist psychiatric services often begin with the assessment of clinical and psychosocial needs of patients by healthcare professionals. Particularly for patients with complex needs or severe problems, ratings of clinical and social functioning at the start of each episode of care may serve as a baseline against which subsequent measures can be compared. Currently, little is known about service variations in such assessments on referrals from primary care. We set out to quantify variability in initial assessments performed by healthcare professionals in three CMHTs in Bristol (UK) using the Health of the Nation Outcome Scales (HoNOS). We tested the hypothesis that variations in HoNOS total and sub-scale scores are related to referral source (general practices), healthcare assessor (in CMHTs) and the assessor's professional group. Statistical analysis was performed using multilevel variance components models with cross-classified random effects. We found that variation due to assessor substantially exceeded that due to referral source (general practices). Furthermore, patient variance differed by assessor profession for the HoNOS – Impairment scores. Assessor variance differed by assessor profession for the HoNOS – Social scores. As HoNOS total and subscale scores show much larger variation by assessor than by referral source, investigations of HoNOS scores *must* take assessors into account. Services should implement and evaluate interdisciplinary training to improve consistency in use of rating thresholds; such initiatives could be evaluated using these extensions of multilevel models. Future research should aim to integrate routine diagnostic data with continuous outcomes to address selection effects (of patients to assessors) better.

Introduction

Episodes of specialist mental healthcare with secondary psychiatric services often begin with an assessment of the range and severity of clinical and psychosocial needs of patients by healthcare professionals (Phelan et al., 2001). In the UK National Health Service (NHS) referrals from primary care general practices must be assessed promptly and comprehensively by members of community mental health teams (CMHTs). Community mental health teams are teams of mental health workers comprising community psychiatric nurses, social workers, psychologists, psychiatrists, occupational therapists (OTs), care workers and administrative staff. The team carries out initial assessments and identifies

people who can provide support and appropriate treatment either in the community or using residential facilities for acute psychiatric care (including hospital beds). Specialist psychiatric services in the UK NHS are only accessed via referral from a general practitioner to the CMHT, except in emergencies. Community mental health teams may provide a range of community-based services themselves but may be complemented by one or more teams providing specific functions, for example for drug and alcohol disorders. Intake assessments of consecutive referrals to CMHTs usually occur after some preliminary triage process in which inappropriate referrals (such as those outside of catchment area) are filtered out: then patients are allocated to healthcare professionals

within each CMHT, according to urgency/availability, need or the skills of staff members. Currently little is known about service variations in such assessments. Furthermore, data arising from *initial* face-to-face contact between patients and mental health professionals are rarely presented or discussed (Evans et al., 2002). This is surprising because intake assessments, at the start of each new episode of care must routinely be conducted on all referrals that have complex needs requiring the involvement of more than one healthcare professional, or more severe or long term illness.

In the UK, guidance issued by the Royal College of Psychiatrists (RCPsych, 2000) recommends that CMHTs take their referrals through a single point of entry from general practitioners (GPs) and requires that all referrals should be assessed promptly by adequately trained staff, with psychiatrists involved in the majority of new-patient assessments. Guidance regarding care programming included in the National Service Framework for Mental Health (Department of Health, 1999) recommends that patients on the highest (second) level of the Care Programme Approach (called 'enhanced CPA') intended for patients with more complex or severe needs) are assessed systematically at initial and subsequent care reviews (Amin et al., 1999).

Clearly there would be benefits if these assessments were standardized, and clinical assessors trained to apply similar thresholds to ratings of psychopathology and social functioning (Huxley et al., 2000). Current practice within secondary care often falls short of such aspirations, and compounds earlier deficiencies in the communication of information between primary and secondary care healthcare professionals (Blaney and Pullen, 1989; Burbach and Harding, 1997).

In this paper we present, for the first time, a novel analysis of data from this key primary-secondary care interface in the UK NHS. We describe some hierarchical aspects of clinical and psychosocial ratings from intake assessments on a large sample of consecutive referrals to psychiatric services in Bristol (Avon, UK) and demonstrate how recently developed statistical models – multilevel models for cross-classified structures – can be used to evaluate hypotheses concerning these variations.

Our analyses capitalize on the fact that secondary care assessors and referring practices overlap, with more than one healthcare professional in the team assessing referrals from any given practice and each assessor often assessing from more than one practice (Figure 1).

This situation enables variability in ratings due to assessor to be statistically disentangled from that due to practice. Investigations of differences in the means and variances of clinical ratings at this interface should be of considerable interest because they identify whether practices refer similar types of patients (the casemix of referrals), whether secondary care assessors receive similar patients, and whether there are differences between professional groups in average ratings or variability. We apply these multilevel extensions of traditional variance components analysis to total and subscale scores on the Health of the Nation Outcome (HoNOS) Scales (Wing et al., 1998; Wing et al., 2000). Assessment of variation between assessors and referring practices is essential to the formulation of appropriate models for change in HoNOS scores during the subsequent spell of care: ignoring this aspect of routine ratings will result in overly precise statements about change over time (improvement, deterioration) or stability of scores.

The HoNOS were developed and are recommended for routine use in the UK as an outcome measure for mental health services (Wing et al., 1998; Bebbington et al., 1999; Wing et al., 2000). They have been widely implemented in the Commonwealth (Trauer et al., 1999; Brooks, 2000; Preston, 2000) and some European countries (Bech et al., 2003). They were designed to measure the inputs and outcomes of mental health services (Wing et al., 1996) but can also help understand the processes of care. In the UK HoNOS has been accepted by the Committee for Regulating Information Requirements for entry in the NHS Data Dictionary. The scales form a key part of the revised Minimum Data Set for Mental Health (Glover, 2000) in the UK.

It is therefore of considerable interest to study patterns of variability in HoNOS scores at the primary secondary care interface, and to relate these to professional groups. Existing reports have primarily involved cross-sectional data or two time-point (entry-discharge) assessments where repeated ratings have been performed. No study to date has emphasized hierarchical (nesting and cross-classification) aspects of the clinical rating context in healthcare organizations. Here we explore these features for the first time using multilevel models.

Methods

Service, sample and setting

The study sample comprised consecutive adult referrals

(age 16–64 years) to secondary care psychiatric services in Avon (Bristol), UK. Data were collected between June and November 1997. Three of the four geographical sectors of Central and South Bristol participated. The catchment area comprised a mix of urban and rural areas with a range of social deprivation (Croudace et al., 2003). The mental health services receiving referrals comprised three generic community mental health teams and three specialist teams: psychopharmacology (academic unit), an alcohol treatment team and a service for patients with drug dependence. Analyses were restricted to the CMHTs only since the drug and alcohol services had more extreme (high > 50%) did-not-attend (DNA) rates (see 'data collection' below). Patients referred to the psychopharmacology service were excluded because it is primarily an academic unit supporting research investigations.

Assessment procedure

The secondary care assessor who is assigned to each referral is required (under guidelines for care planning) to perform a thorough evaluation of the patient and to develop a care plan. Increasingly it is expected that, for patients with complex needs or severe illness, this assessment include some formal (structured) appraisal of need for care, or a clinical rating of psychopathology and psychosocial functioning, which captures the problems to be addressed.

These assessment processes result in the formation of a care plan, and generate formal documentation that must be reviewed by other healthcare professionals participating in care provision.

Clinical assessors (healthcare professional working in secondary mental healthcare)

Professional groups employed within the secondary services studied were psychiatrists ($n = 28$), occupational therapists (OTs) ($n = 7$), psychologists and counsellors ($n = 6$), community psychiatric nurses ($n = 22$) assessing respectively 138 (36%), 29 (8%), 51 (13%) and 166 (43%) patients. All groups participated in provision of clinical ratings.

Assessment instrument: the Health of the Nation Outcome Scales

All healthcare professionals in the CMHTs and specialist services were asked to complete ratings on each

referral using the Health of the Nation Outcome scales (see Appendix 1; Wing et al., 1999; and also the Web sites at <http://www.rcpsych.ac.uk/cru/honoscales/faqs.htm> and <http://www.rcpsych.ac.uk/cru/honoscales>). Briefly, the HoNOS comprise 12 items that consider different aspects of mental health and social functioning. Each item is rated on a five-point scale that is scored either zero through four or one through five (the former used here). A full glossary (<http://www.rcpsych.ac.uk/cru/honoscales>) and instructions for use are distributed with the instrument.

It has been suggested that HoNOS scores, from such assessments (and subsequent care reviews) can contribute useful (baseline) measures of clinical or social functioning that may serve as a reference point against which subsequent assessments can be compared, thereby offering a perspective on areas of improvement or recovery. Repeating assessments at subsequent care reviews could give rise to extensive longitudinal data. In this study we only consider a single assessment since no longitudinal data (repeat assessment) were available.

Training

Assessors were exposed to vignette-based training materials, but did not receive all elements of training, for example videos, now proposed by the developers.

Data collection

The HoNOS forms were posted to clinical assessors by a research assistant who was informed of the outcome of the community mental health team's allocation (triage) meetings. Ratings were sustained for 6 months with some administrative reminders by phone or letter performed by the research assistant. In routine psychiatric care non-attendance is a huge problem with a significant proportion (between 20% and 50%) of referrals not attending their appointments (Evans et al., 2002). In such circumstances where no face-to-face contact was made there was no opportunity for clinical ratings to have been made. We do not therefore consider referrals that did not attend their appointment with the CMHT but know that they were more likely to be younger patients (aged <35 years). The DNA rates did not differ between men and women. Since DNA rates for drug and alcohol teams approached 50% we only considered referrals to the CMHTs for whom DNA rates were lower (by approximately 25%) in keeping with anecdotal reports from other UK

services. Further details on the sample are reported in Evans et al. (2002).

Multilevel variance components models: justification and definition

At the primary-secondary care interface between general practices and specialist mental health services, referring practices (primary care) and assessor (health-care professional in secondary care) are arranged in a cross-classified hierarchy: they are not simply nested. This feature of organisational overlap is illustrated schematically in the crossing lines of Figure 1.

Investigations of differences in the means and variances of clinical ratings at this interface should be of considerable interest because they identify whether practices refer similar types of patients (the casemix of referrals), whether secondary care assessors receive similar patients, and whether there are differences between professional groups in average ratings, or variability. Testing such hypotheses requires an extension of traditional variance components models to multilevel structures with non-nested (cross-classified) units. Such models were formulated in an educational context by Rasbash and Goldstein (1994) and Raudenbush (1993), and have been applied in healthcare research by Leyland and Goldstein (2001). This paper represents the first attempt to apply these models to routine outcome data on psychiatric referrals.

Aim of the multilevel analysis

The goal of our analysis was to bring a multilevel variance components perspective (Raudenbush, 1993; Snijders and Bosker, 1999; Goldstein, 2002) to modelling of HoNOS scores on a sample of consecutive referrals. In our model, individual variability (the HoNOS score for each referral) is assumed to be influenced by (at least) three factors: the referring practice (where), the secondary care assessor (who), and their (the assessor's) professional group. Our model is designed so as to quantify the contribution of these contextual factors and individual characteristics to differences in both level and variability of HoNOS (total and subscale) scores.

In formal notation multilevel models for such cross-classified structure can be represented as:

$$y_{i(p,a)} = X_{i(p,a)} \beta + v_p + u_a + e_{i(p,a)} \quad (1)$$

$$p = 1, \dots, P, a = 1, \dots, A, i = 1, \dots, N(p, a)$$

in which the score, $y_{i(p,a)}$, of the patient i (level 1), nested within a combination of practice (p) and assessor (a) (level 2), is predicted by explanatory variables $X_{i(p,a)}$ and their associated coefficients β . We note that such models can, in theory, include patient, practice and assessor covariates, but only assessor covariates are considered here. A justification for introduction of more (patient level) covariates (X variables describing the characteristics of referrals) is presented in the

Figure 1: Level (in Multilevel model): note there are two higher levels (2s).

discussion. Within such a model it is possible to define two intra-class correlations of interest:

- for referrals nested within practices, and
- for referrals nested within assessors

using standard formulae for random effects ANOVA (Goldstein, 2002).

Fixed effects

The fixed effects in all our models are dummy variables for assessor profession that compare psychiatrists, psychologists and counsellors, and OTs, with community psychiatric nurses (the modal and reference category).

Random effects

The random ($v_p + u_a + e_{i(p,a)}$) part of the model is given by two (level 2) random effect terms, one for practice (v_p) and one for assessor (u_a), together with the usual (individual level 1) residual ($e_{i(p,a)}$). The variance of the assessor random effects and the individual residuals (u_a and $e_{i(p,a)}$) were allowed to vary by assessor profession after preliminary analysis identified complex variance structures at levels 1 and 2. Need for complex level 1 and level 2 variance structures was evaluated using likelihood ratio (chi-square) tests with 3 degrees of freedom. These features of the model are illustrated schematically, in Figure 2, in order to clarify interpretation.

Here we illustrate different types of variation by assessor. In Figure 2 each box plot shows the range of HoNOS scores for patients rated by one assessor. In each panel, the three left-most box plots represent three assessors from the same professional group: they have slightly different means, indicating $\text{var}(u_a) > 0$. The three right-most box plots in each panel represent three assessors from a second professional group. In panel A, the individual level variance (level 1 residual), $\text{var}(e_{i(p,a)})$, is larger in the second group, as each assessor returns a wider range of ratings. In panel B, the level 2 assessor variance, $\text{var}(u_a)$, is larger in the second group, as the assessor averages differ more widely. Panel C shows a more familiar situation where both variances are equal but the overall mean is larger in the second group. This effect is summarized by a simple fixed effect indicating the difference in group means. In any model (for each HoNOS outcome) one or more of these features may apply simultaneously, giving rise to realistic complexity in the modelling of cross-classified data.

Figure 2 Schematic representation of different types of variability at different levels in a multilevel model.

Each box plot shows the range of scores for patients rated by a rater. The 3 left-most box plots represent 3 raters from the same professional group. The 3 rightmost box plots in each panel represent three raters from a second professional group. Interest focuses on the similarities in the means (level 2 variation) and variances (level 1).

Health of the Nation Outcome Scales variables: distribution and transformation

The outcome variables were the HoNOS total score (sum of all items) and four subscale scores, computed as the sum of items for each section (1. Behaviour, 2. Impairment, 3. Symptoms, 4. Social; for component items see Appendix 1).

Total and subscale scores exhibited considerable positive skewness, so all scores were transformed prior to the analysis by assigning values from a standard normal distribution according to the ranks of the

scores in the sample. This was very effective for the total score, but less so for the subscales (See Figure 3b).

Model estimation

The method of Rasbash and Goldstein (1994) for estimating cross-classified models was used. This converts the model into a simple hierarchical model by considering all units of one of the cross-classified random variables as a set of dummy variables (0,1) introduced as explanatory variables into the random part of the model with variances constrained to be equal: this provides an estimate of the between unit variance.

All models were estimated in MLwiN software (Version 1.20) (Rasbash et al., 2000) using restricted maximum likelihood. Identical results were returned using the MIXED procedure in SAS (SAS Institute, 1989). Model syntax and full output are available from TJC.

Hypothesis tests

Fixed effects were tested both using Wald tests (in

MLwiN) and by multiple (three) degree of freedom F tests with correction to denominator degrees of freedom (available in SAS). Conclusions did not differ. Presence of random variances, and equality of random variances between professional groups, was tested by likelihood ratio tests of nested models. All models for likelihood ratio comparisons were estimated using maximum likelihood (ML), but final models were estimated using restricted maximum likelihood (REML) so as not to underestimate random effect variances. Tests of fixed effects were performed in models with the most complex variance structures indicated after appropriate likelihood ratio tests for random effects.

Results

Descriptive statistics

Table 1 shows the frequency distribution of responses to the twelve HoNOS items (H1-H12) for the analysis sample which comprises 384 patients, referred to south Bristol CMHTs over a 6-month period, who both

Table 1. Distribution of responses for HoNOS items: number (%), n = 384 with complete data on all HoNOS items

HoNOS Item	0: No problem	1: Minor problem	2: Mild problem	3: Moderately severe problem	4: Severe to very severe problem
H1 Overactive, aggressive, disruptive or agitated behaviour	250 (65.1%)	76 (19.8%)	40 (10.4%)	13 (3.4%)	5 (1.3%)
H2 Non-accidental self-injury	290 (75.5%)	52 (13.5%)	19 (4.9%)	12 (3.1%)	11 (2.9%)
H3 Problem drinking	255 (66.4%)	37 (9.6%)	38 (9.9%)	35 (9.1%)	19 (4.9%)
H4 Cognitive problems	256 (66.7%)	78 (20.3%)	36 (9.4%)	12 (3.1%)	2 (0.5%)
H5 Physical illness or	282 (73.4%)	41 (10.7%)	32 (8.3%)	21 (5.5%)	8 (2.1%)
H6 Problems with hallucinations and delusions	349 (90.9%)	10 (2.6%)	13 (3.4%)	8 (2.1%)	4 (1.0%)
H7 Problems with depressed mood	83 (21.6%)	65 (16.9%)	130 (33.9%)	92 (24.0%)	14 (3.6%)
H8 Other mental and behavioural problems	174 (45.3%)	19 (4.9%)	76 (19.8%)	90 (23.4%)	25 (6.5%)
H9 Problems with relationships	135 (35.2%)	92 (24.0%)	66 (17.2%)	76 (19.8%)	15 (3.9%)
H10 Problems with activities of daily living	228 (59.4%)	77 (20.1%)	49 (12.8%)	22 (5.7%)	8 (2.1%)
H11 Problems with living conditions	292 (76.0%)	56 (14.6%)	19 (4.9%)	11 (2.9%)	6 (1.6%)
H12 Problems with occupational and activities	258 (67.2%)	60 (15.6)	34 (8.9%)	24 (6.3%)	8 (2.1%)

attended their appointment and had complete data on all HoNOS items (rated by a staff member). Sixty-three different assessors contributed clinical ratings.

For all items with the exception of problems with depressed mood the modal response was the first category, absence of any problem. Items with the highest prevalence of responses in the moderate or severe categories (the two highest ratings) included drinking or drug taking (H3), problems with depressed mood (H7), other mental and behavioural problems (H8),

and problems with relationships (H9).

Figures 3a and 3b show the distribution of the HoNOS total and subscale scores respectively (all after transformation). Clearly transformation achieved an approximate normal distribution for the total scores and reduced skew for the subscales, but some deviation from normality remained.

Table 2 reports all parameters (estimates and standard errors) from final models for the total score, before and after transformation to normality, together

Figure 3a. Histograms for HoNOS total scores (transformed to normality). Overlay = standard normal distribution.

Figure 3b. Histograms for HoNOS subscale scores (transformed to normality).

with tests of zero variance for each level 2 random effect (practice and assessor) and for equality of random effect variances across assessors.

Between-practice and between-assessor components of variance

Health of the Nation Outcome Scales total score

Of the total variability in the non-transformed HoNOS score, 4.5% (intra-class correlation, 0.045) was between practices and 28.1% (intra-class correlation, 0.281) between assessors. The remaining variability (67.4%; the largest component of variance) was between individual patients (referrals). The variance between assessors was highly statistically significant. In contrast the variance between practices

was small. There was no evidence that profession of assessor affected either the variance (complex level 2 variance structure) or the variance between patients (complex level 1 variance structure). The coefficients for the fixed effects, reported in the lower half of Table 2, indicate that there were no differences in mean scores between assessor professional groups ($\chi^2_3 = 2.27$, $p = 0.52$). The residuals at both levels 1 and 2 (not shown) were found to be reasonably normally distributed at both levels 1 and 2 (standardized skewness both <1.0). Results for the transformed total score agreed with those for the non-transformed total score.

Health of the Nation Outcome Scales subscale scores

All HoNOS subscales show variability due to assessor but none showed significant variability due to practice

Table 2. Cross-classified multilevel model for HoNOS total score $n = 384$

	Total score			Total score (transformed)		
	Estimate (standard error)	Zero variance (p-value)	Equal between assessor professions (p-value)	Estimate (standard error)	Zero variance (p-value)	Equal between assessor professions (p-value)
<i>Random effects</i>						
Level 2 – practice	1.524 (1.049)	0.34	0.48	0.000 (0.000)	1.00	0.97
Level 2 – assessor		<0.01			<0.01	
Psychologist	9.445 (2.616)			0.261 (0.072)		
Psychiatrist	9.445 (2.616)			0.261 (0.072)		
Occ. therapist	9.445 (2.616)			0.261 (0.072)		
Nurse	9.445 (2.616)			0.261 (0.072)		
Level 1-patient			0.27			0.24
Psychologist	22.680 (1.856)			0.683 (0.0533)		
Psychiatrist	22.680 (1.856)			0.683 (0.0533)		
Occ. therapist	22.680 (1.856)			0.683 (0.0533)		
Nurse	22.680 (1.856)			0.683 (0.0533)		
<i>Fixed effects</i>						
Constant (nurse)	9.58 (0.83)			0.06 (0.13)		
Psychiatrist	−2.19 (1.76)			−0.31 (0.29)		
Occ. ther. nurse	−1.34 (1.10)			−0.21 (0.18)		
Psychologist nurse	−0.96 (1.67)			−0.18 (0.28)		
<i>Intra-class correlations</i>						
Level 2 – practice	4.5			0.0		
Level 2 – assessor	28.1			29.1		

Note: Intra-class correlation estimates reported here, and in subsequent tables, are from REML estimates of model with fixed effects and both practice and assessor cross-classified random effects (no complex level 1 or two variance structures). Practice variance reported here is also estimated from the same model, but tested using likelihood ratio difference test between models with and without practice, estimated using maximum likelihood.

(Table 3). For subscales 1 and 3 (1 – Behaviour, 3 – Symptoms) neither the variance between assessors nor that between patients showed any variation according to assessor profession. For subscale 2 (HoNOS – Impairment), the variance between patients differed between assessor professions (see panel A of Figure 2), highest variability being found for OTs and nurses. This is an example of complex level 1 variance structure. The intra-assessor correlation estimates varied between 0.314 for psychiatrists, 0.256 for psychologists, 0.198 for nurses and 0.197 for OTs.

For subscale 4 (HoNOS – Social), the variance between assessors differed between assessor professions (see panel B of Figure 2). Variability was found for nurses and psychiatrists but not for psychologists and occupational therapists. The intra-assessor correlation estimates varied between 0.373 for nurse, 0.300 for psychiatrist and 0.00 for psychologist and occupational therapists.

None of the subscales showed any difference in mean values for the different assessor professions – there were no main effects on average ratings between occupation groups.

Discussion

This paper represents the first attempt to model the determinants of service variations at the primary-secondary care interface for psychiatric referrals to UK mental health services. We applied hierarchical (multilevel) models appropriate for hierarchically cross-classified structures to model practice and assessor variation in HoNOS scores from intake assessments.

Our goal was to model the (random) effects of practice and assessor and so to enable them to be separated statistically. In doing so we determined both the magnitude and significance of these cross-classified variance components and explored differences between average

Table 3. Cross-classified multilevel model for HoNOS subscale scores (transformed)

	HoNOS Subscale 1 Behaviour			HoNOS Subscale 2 Impairment		
	Estimate (standard error)	Zero variance (p-value)	Equal between assessor professions (p-value)	Estimate (standard error)	Zero variance (p-value)	Equal between assessor professions (p-value)
<i>Random effects</i>						
Level 2 – practice	0.0	1.00		0.004 (0.020)	1.00	
Level 2 – assessor		<0.01	0.49		<0.01	0.91
Nurse	0.165 (0.051)			0.155 (0.043)		
Psychiatrist	0.165 (0.051)			0.155 (0.043)		
Occ. therapist	0.165 (0.051)			0.155 (0.043)		
Psychologist	0.165 (0.051)			0.155 (0.043)		
Level 1-patient			0.45			0.004
Nurse	0.597 (0.047)			0.622 (0.072)		
Psychiatrist	0.597 (0.047)			0.335 (0.044)		
Occ. therapist	0.597 (0.047)			0.628 (0.178)		
Psychologist	0.597 (0.047)			0.446 (0.093)		
<i>Fixed effects</i>						
Constant (nurse)	0.104 (0.112)			0.149 (0.110)		
Psychiatrist	–0.029 (0.155)			–0.222 (0.248)		
Occ. ther. nurse	–0.353 (0.250)			–0.304 (0.145)		
Psychologist nurse	–0.276 (0.232)			–0.067 (0.220)		
<i>Intra-class correlations</i>						
Level 2 – practice	0.0			0.5/0.8/0.5/0.7		
Level 2 – assessor	21.7			19.7/31.4/19.8/25.6		

ratings between professional groups.

We expected to demonstrate that referral source (general practices), referrals (the patients) and assessors (the healthcare professionals in the CMHTs) would each contribute systematically to variability in HoNOS ratings. In fact, we found little or no variability in HoNOS scores due to the source of referral – the general practices. This is intuitively reasonable and important as it confirms that the referral source can be ignored in multilevel analyses if the secondary care assessor is taken into account. This greatly simplifies modelling procedures.

However we caution that if secondary care assessors are ignored and so not incorporated into the multilevel model, then their impact will be haphazardly distributed among the referring practices. Clearly it is better to attribute the variance to the most likely source – the assessors in our case. These results allows

us to recommend that relatively simple multilevel models be applied to referrals nested within assessors, without the need for further consideration of more complex multilevel models with cross-classified structures of the sort used here. These models would be particularly important for assessment of change over secondary care episodes when a different rater assesses the patient at entry and exit from care. The cross-classified models used in this paper would be necessary if the raters differ and independent effects of rater are found on both occasions. Failure to acknowledge this variability between assessors in models of change will result in unwarranted precision of estimates of change.

We note that we did not explore differences due to referring general practitioners, only general practices. A different set of hypotheses may motivate interest in individual general practitioners.

Variability was found between assessors (within

Table 3. Cont.

	HoNOS Subscale 3 Symptoms			HoNOS Subscale 4 Social		
	Estimate (standard error)	Zero variance (p-value)	Equal between assessor professions (p-value)	Estimate (standard error)	Zero variance (p-value)	Equal between assessor professions (p-value)
<i>Random effects</i>						
Level 2 – practice	0.026 (0.036)	0.08		0.000 (0.000)	1.00	
Level 2 – assessor		<0.01	0.10		<0.01	0.05
Nurse	0.138 (0.056)			0.335 (0.133)		
Psychiatrist	0.138 (0.056)			0.241 (0.102)		
Occ. therapist	0.138 (0.056)			0.000 (0.000)		
Psychologist	0.138 (0.056)			0.000 (0.000)		
Level 1-patient			0.14			0.09
Nurse	0.752 (0.060)			0.520 (0.043)		
Psychiatrist	0.752 (0.060)			0.520 (0.043)		
Occ. therapist	0.752 (0.060)			0.520 (0.043)		
Psychologist	0.752 (0.060)			0.520 (0.043)		
<i>Fixed effects</i>						
Constant (nurse)	–0.033 (0.116)			0.101 (0.143)		
Psychiatrist	–0.340 (0.260)			–0.425 (0.200)		
Occ. ther. nurse	–0.032 (0.160)			–0.230 (0.187)		
Psychologist nurse	–0.124 (0.234)			–0.145 (0.178)		
<i>Intra-class correlations</i>						
Level 2 – practice	2.8			0.0		
Level 2 – assessor	15.0			37.3/30.0/0.0/0.0		

practices) for only one of the subscales (subscale 4 – social) but not for total scores. The intra-class correlation at the assessor level for (transformed) total score was 0.29, and ranged from 0.00 to 0.37 for the four subscales.

Importantly, there were no differences in average ratings between referrals assessed by the different professional groups. This result is captured in models by standard dummy variables.

Of course, by far the largest component of variability (67% for the total score) was between individual referrals, confirming the wide variation in severity and need among referrals. Although not altogether surprising, it is pleasing that our model could quantitatively confirm and summarize the substantially larger variability found here, than that evident between assessors.

It is important to emphasize that this was a naturalistic study, so we cannot directly conclude that variation in ratings between assessors represents variation in the rating behaviours of these groups, because the results may be influenced by effects of selection of patients to assessor profession and to assessors within assessor profession. For example, variance between assessors would arise if patients with greater need, severity or complexity of problems were assigned to specific assessors, perhaps because of their skills and experience.

It is entirely proper that referrals are allocated on the basis of fit between the characteristics of the referral and the skills offered by the CMHT staff members who will perform the assessment and provide or co-ordinate the care plan. The greater variability in the social subscale score between nurses and psychiatrists than between occupational therapists and psychologists could be due to selection (of particular diagnoses) to nurses and psychiatrists, this could be examined in the context of concurrent data on routine diagnoses (a fruitful area for future research). Alternatively there may be factors attributable to particular assessors or professions that determine the variability of ratings for a given assessor profession. We note that Amin et al. (1999) found weakest correlations between key-worker and researcher ratings on the social subscale, suggesting possible problems with the interpretation of these items. Elsewhere it has been recommended that items 11 and 12 be dropped from the total score. There may therefore be particular issues to do with the thresholds defined for rating these items, or with the

information available at the time to the clinical assessors.

Other limitations include the high proportion of non-attendance, which may have caused bias. The HoNOS ratings were collected during a survey of referrals prior to an intervention study, but in all other senses the rating context was akin to normal clinical care. We believe that the Bristol service is probably similar to other urban psychiatric services in terms of the factors contributing to non-attendance, with young patients particularly likely not to attend.

Finally, HoNOS items were summed assuming continuous scores although data are strictly ordinal. This may have contributed to the only approximate normality of some of the subscale scores, even after transformation. Our outcome comprised HoNOS scores at the interface between primary and secondary mental health services, but our methods can be applied to any continuous rating resulting from clinical or needs assessment protocols. Extensions of our multi-level modelling framework to ordinal outcomes allow similar models to be estimated for individual HoNOS items. Incorporation of latent variables, in the form of common factors, offer an alternative approach to errors of measurement.

These themes may be integrated through multilevel factor analysis models or hierarchical extensions of item response theory methods. We encourage these developments in future work combining latent factors and multilevel models for hierarchical data. Within such a comprehensive modelling framework it becomes a simple matter to extend this modelling approach to examine the effect of structured interventions, such as training, to standardize HoNOS ratings, or reduce between individual or professional group variability. Threshold inconsistency may be amenable to intervention (cf. Rock and Preston, 2001) and appropriate training may help in standardizing ratings (Muller and Wetzel, 1998) through use of common thresholds.

A specific concern for training is suggested by the relatively high variability of HoNOS ratings within the nurse and psychiatrist groups for the social subscale. Such training would have the aim of increasing consistency across assessors in use of thresholds for rating HoNOS items, and indirectly improving the comparability of subscale and total scores. When repeated (annually) this may prevent ‘assessor drift’. Initially this approach might be applied in a pre-post

study of training for a subset of assessors, to examine the impact of training (Rock and Preston, 2001) on the relevant variance components (patients and assessors).

In conclusion, we have demonstrated that the effect on ratings of the professional group of secondary care assessors cannot be ignored, because they have a major influence on variability, but that the role of referral source appears minimal (and can be ignored). We believe that only part of the observed effect of assessors is likely to be due to selection, hence our recommendation of training interventions for assessors to increase consistency in use of scale scores. These claims need to be further defended in the context of more comprehensive adjustments for additional patient and professional factors (for example experience, Roncone et al., 1999). However, one clear implication is that assessment training should take into account professional group, most practically by a multidisciplinary approach.

Acknowledgements

Funding for this secondary analysis was provided by a project grant from the former NHS Eastern Regional NHS Executive, now known as the Department of Health, Directorate of Health and Social Care (Midlands and East of England): *Psychometric evaluation of the revised NHS Mental Health Minimum Data*. Grant Number HSR/0301/2 (Reference: RCC33034). Original data collection was supported by a project grant from the UK Department of Health (grant reference JRI26/0634).

References

- Amin S, Singh SP, Croudace T, Medley I, Harrison G. Evaluating the Health of the Nation Outcome Scales: reliability and validity in a three-year follow-up of first-onset psychosis. *British Journal of Psychiatry* 1999; 174: 399–403.
- Bebbington P, Brugha T, Hill T, Marsden L, Window S (1999) Validation of the Health of the Nation Outcome Scales, *British Journal of Psychiatry* 1999; 174: 389–94.
- Bech P, Bille J, Schutze T, Sondergaard S, Wiese M, Waerst S. Health of the Nation Outcome Scales (HoNOS): implementability, subscale structure and responsiveness in the daily psychiatric hospital routine over the first 18 months. *Nordic Journal of Psychiatry* 2003; 57, 285–90.
- Blaney D, Pullen I. Communication between psychiatrists and general practitioners: what style of letters do psychiatrists prefer? *Journal of the Royal College of General Practitioners* 1989; 39: 67.
- Brooks, R (2000) The reliability and validity of the Health of the Nation Outcome Scales: validation in relation to patient derived measures, *Australian and New Zealand Journal of Psychiatry* 2000; 34(3): 504–11.
- Burbach FR and Harding S. GP referral letters to a community mental health team: an analysis of the quality and quantity of information. *International Journal of Health Care Quality Assurance* 1997; 10: 67–72.
- Croudace T, Evans J, Harrison G, Sharp DJ, Wilkinson E, McCann G, Spence M, Crilly C, Brindle L. Impact of the ICD-10 Primary Health Care (PHC) diagnostic and management guidelines for mental disorders on detection and outcome in primary care: cluster randomised controlled trial. *British Journal of Psychiatry* 2003; 82: 20–30.
- Department of Health. National Service Framework for Mental Health. London: Stationery Office; 1999.
- Evans J, Wilkinson E, Brindle L, Harrison G, Sharp D, Croudace T, McCann G. Clinician opinions about the appropriateness and severity of general practitioner referrals to specialist mental health services: a cross-sectional survey. *Primary Care Psychiatry* 2002; 8(3): 91–4.
- Glover GR. A comprehensive clinical database for mental healthcare in England. *Social Psychiatry and Psychiatric Epidemiology* 2000; 35(11): 523–9.
- Goldstein H. *Multilevel Statistical Models*. (Third Edition). London: Arnold, 2002.
- Huxley P, Reilly S, Gater R, Robinshaw E, Harrison J, Mohamed H, Butler T, Windle B. Matching resources to care: the acceptability, validity and inter-rater reliability of a new instrument to assess severe mental illness (MARC-1). *Social Psychiatry and Psychiatric Epidemiology* 2000; 35: 312–7.
- Leyland A Goldstein H. *Multilevel Modelling of Health Statistics*, Chichester, Wiley, 2001.
- Muller MJ, Wetzel H. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. *Acta Psychiatrica Scandinavica* 1998; 98(2): 135–9.
- Preston NJ. The Health of the Nation Outcome Scales: validating factorial structure and invariance across two health services. *Australian and New Zealand Journal of Psychiatry* 2000; 34: 512–9.
- Phelan M, Sellar J, Leese M. The routine assessment of severity amongst people with mental illness. *Social Psychiatry and Psychiatric Epidemiology* 2001; 36: 200–6.
- Rasbash J, Goldstein H. Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics* 1994; 19: 337–50.
- Rasbash J, Browne W, Goldstein H, Yang M, Plewts I, Healy M, Woodhouse G, Draper D, Hangford I, Lewis T. *An User's Guide to MLwiN*. Multilevel Models Project, Institute of Education, University of London, 2000.

- Raudenbush SW (1993) A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics* 18: 321–49.
- Rock D, Preston N. HoNOS: is there any point in training clinicians? *Journal of Psychiatric and Mental Health Nursing* 2001; 8: 405–9.
- Roncione R, Ventura J, Impallomeni M, Falloon IHR, Morosini PL, Chiaravalle E, Casacchia M. Reliability of an Italian standardized and expanded Brief Psychiatric Rating Scale (BPRS 4.0) in raters with high vs. low clinical experience. *Acta Psychiatrica Scandinavica* 1999; 100(3): 229–36.
- SAS Institute, Inc. SAS/STAT Software: Changes and Enhancements, Release 8.1. Cary NC, 2000.
- Snijders TAB, Bosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage, 1999.
- Trauer T, Callaly P, Hantz P, Little J, Shields R, Smith J. Health of the Nation Outcome Scales. Results of the Victorian field trial. *The British Journal of Psychiatry* 1999; 174: 380–8.
- Wing JK, Curtis RH, Beevor AS. HoNOS: Health of the Nation Outcome Scales: Report on Research and Development July 1993 – December 1995. London: Royal College of Psychiatrists, 1996.
- Wing JK, Beevor AS, Curtis RH, Park SBG, Hadden S, Burns A. Health of the Nation Outcome Scales (HoNOS): Research and development. *British Journal of Psychiatry* 1998; 172, 11–18.
- Wing JK, Curtis RH, Beevor AS. Health of the Nation Outcome Scales (HoNOS): Glossary for HoNOS score sheet. *British Journal of Psychiatry* 1999; 174: 432–4.
- Wing JK, Lelliott P, Beevor AS. Progress on HoNOS. *British Journal of Psychiatry* 2000; 176, 392–3.

Appendix 1: The Health of the Nation Outcome Scales (Wing et al., 1996, 1999)

HoNOS – Behaviour (Subscale 1 H1–H3)

- H1 Overactive, aggressive, disruptive or agitated behaviour
- H2 Non-accidental self-injury
- H3 Problem drinking or drug-taking

HoNOS – Impairment (Subscale 2 H4–H5)

- H4 Cognitive problems
- H5 Physical illness or disability problems

HoNOS – Symptoms (Subscale 3 H6–H8)

- H6 Problems associated with hallucinations and delusions
- H7 Problems with depressed mood
- H8 Other mental and behavioural problems

HoNOS – Social (Subscale 4 H9–H12)

- H9 Problems with relationships
- H10 Problems with activities of daily living
- H11 Problems with living conditions
- H12 Problems with occupation and activities

Responses to all items are made on five-point scales:

- 0 = No problem
- 1 = Minor problem requiring no action
- 2 = Mild problem but definitely present
- 3 = Moderately severe problem
- 4 = Severe to very severe problem

Correspondence: TJ Croudace, Department of Psychiatry, University of Cambridge, Box 189, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK.
 Telephone (+44) 1223 336599.
 Fax (+44) 1223 336968.
 Email tjc39@cam.ac.uk.