Comprehensive Psychiatry 93 (2019) 61–64

Contents lists available at ScienceDirect

Comprehensive Psychiatry

journal homepage: www.elsevier.com/locate/comppsych

A model for predicting missing items on the Health of the Nation Outcome Scale (HoNOS)

Tim M. Gale *, Billy Boland

Hertfordshire Partnership University NHS Foundation Trust, UK
University of Hertfordshire, UK

ARTICLE INFO

Keywords:
HoNOS
Missing data
Outcome measures
Mental health
Weighted mean
Clustering

ABSTRACT

HoNOS is one of the most widely used clinician rated outcome measures in mental health services. A commonly encountered problem is that one or more of the 12 individual HoNOS items is left uncompleted (missing data rates of up to 25% have been reported), which affects the degree to which organisations can rely on the accuracy of historical HoNOS data. In this brief paper we outline a simple statistical method of predicting missing item scores for HoNOS, both in general adult and 65+ populations. The method accounts for the average pattern of responding being non-uniform across items: i.e., some HoNOS items consistently elicit higher scores than others. By calculating individual item weights based on a very large sample of fully completed HoNOS assessments, we were able to accurately predict the value of missing items in a new sample. We contrast the accuracy of this approach with two other simple statistical procedures, and show that the weighted means model returns a much lower error rate. Although this is not the only method of predicting missing items, it carries the advantages of being: (i) free of charge, (ii) easily applicable to large datasets using a spreadsheet and (iii) unreliant on the availability of previous assessment data for the same patients. We hope this method will be of use to other organisations that are processing large volumes of HoNOS data.

1. Introduction

HoNOS is one of the most routinely used measures in NHS mental health services. It was developed by [1] during the early 1990s and first published in 1996, with the aim of measuring improvements in the health and social functioning of people with enduring mental health problems [2]. Since its launch, it has been a key outcome indicator for mental wellbeing and, since 2011, its use has increased considerably because of its inclusion in the clustering tool used within the NHS England mental health payment system. Developed by a group in the North East of England [3], 'care clusters' used HoNOS as a foundation to describe needs, with a conscious move away from using a 'purely diagnostic' description of people [4]. Care clusters remain a key part of the National Payment Tariff System in England [5], sometime referred to as 'Payment by Results' for mental health services.

HoNOS is also used in other countries, notably New Zealand and Australia, and distinct versions have now been created for over 65s, children and adolescents, and for people with learning disabilities (see https://www.rcpsych.ac.uk/events/in-house-training/health-of-nation-outcome-scales). Since their inception, the glossaries for scoring the scale had not been revised. Subsequently, the Royal College of Psychiatrists convened an international advisory board to identify 'desirable amendments'. The Royal College Council agreed the proposed changes in 2017, and acknowledged that it was 'desirable' that these revisions be subject to empirical testing for use in the field [6].

Developing outcome measurement as an indication of quality has grown in importance within health services in recent years. HoNOS continues to be championed by the National Collaborating Centre for Mental Health as an outcome measure for a number of care pathways, including first episode psychosis [7] and Dementia [8]. To support care delivery, NHS England and NHS Improvement recommends mental health care providers develop an outcomes framework to monitor and improve quality. HoNOS is proposed as a Clinician Rated Outcome Measure (CROM) to be used within the framework [5,9].

Structurally, HoNOS comprises 12 ratable items (see Table 1), each of which is scored from 0 (no problem), through 1 (minor problem), 2 (mild problem), 3 (moderately severe problem) to 4 (severe to very severe problem). A total score is obtained (ranging from 0 to 48) and subscale scores can also be calculated by combining groups of items as follows: (i) behavioural problems (items 1–3), (ii) impairment (items 4 and 5), (iii) symptomatic problems (items 6–8) and (iv) social problems (items 9–12).

A strength of HoNOS, as a clustering tool and outcome measure, is that it spans a diverse set of constructs, examining not just clinical

* Corresponding author at: Department of Research & Development, HPFT, The Colonnades, Beaconsfield Road, Hatfield AL10 8YE, UK.
E-mail address: t.gale@herts.ac.uk (T.M. Gale).

https://doi.org/10.1016/j.comppsych.2019.06.005

Table 1
The 12 HoNOS items (each rated from 0 to 4, maximum total 48).

Individual HoNOS Items
1. Overactive aggressive, disruptive or agitated behavior
2. Non-accidental self-injury
3. Problem drinking or drug taking
4. Cognitive problems
5. Physical illness or disability problems
6. Problems with hallucinations and delusions
7. Problems with depressed mood
8. Other mental and behavioral problems (specify)
9. Problems with relationships
10. Problems with activities of daily living
11. Problems with living conditions
12. Problems with occupation and activities

features of a patient, but also the impact that mental illness has on daily life. However, the downside of this is that proper completion of the scale necessitates a comprehensive assessment requiring access to information that may not always be readily available. For this reason it is not uncommon for individual items to be left un-scored, with estimates of missing scores for routinely collected HoNOS data being as high as 25% [10].

When a HoNOS item cannot be completed, the standard procedure is to enter the value '9' as an indication for missing data. However this is not helpful beyond a coding level, since it exceeds the maximum item score and has no interpolative value. Missing items cause problems for ongoing monitoring and analyses, making it more difficult to track changes over time. Moreover, for those organisations that might want to look at HoNOS scores as a measure of service quality, a high number of missing items will create noise in the system.

Although complex algorithms exist for dealing with missing data (e.g. [10]), these are not readily available and may require considerable expertise to implement. There are, however, some simple statistical approaches to dealing with missing item data, which might be considered here. For example, a previous study examining two routinely used depression rating scales, The MÅDRS [11] and The BDI [12], compared three simple statistical approaches to estimating missing item values: (i) substitution of missing items with the modal score of all completed items, (ii) substitution with the mean score of all completed items, and (iii) substitution with a weighted mean score of all completed items [13]. Neither the mode nor the mean turned out to be very good predictors for missing item values on these depression scales, for the simple reason that some items consistently elicit a higher score than others. In other words, the pattern of responding for the average depressed patient is not uniform across all items. Under the third approach, a set of weights was derived to capture this pattern, allowing very accurate prediction of missing item values on these two scales.

In this paper, we describe a similar approach applied to HoNOS scores. Given the currently strong emphasis on measuring patient outcomes, we anticipate that an accurate model for estimating missing item scores will be of value to healthcare organisations who routinely collect this data and may wish to make better use of it. A quick and easy means of imputing missing scores may also be of value to clinical teams.

2. Method

2.1. Sample

Our data for this research comprised anonymous HoNOS scores for a cohort of 24,473 adult mental health patients drawn from a secondary care UK mental health trust, over a 5-year period. However, this dataset was reduced down to 22,164 patients after excluding all those patients who had one or more missing item scores (approximately 9% of all records). The fact that nearly one in ten patients from a consecutive clinical sample had missing items, further illustrates the importance of developing an approach to address this issue. We then randomly sampled (without replacement) 500 patients from the dataset of fully intact HoNOS scores to use as a test set for our three predictive models (see later). Of the 21,664 patients remaining, 45% were male and 55% were female. Mean age was 39 (range 18–69, SD 14.8). These patients had been, or were being, treated for a range of mental health conditions within general adult psychiatry and each of these patients contributed two complete HoNOS ratings, a minimum of one month apart. Thus the total number of complete HoNOS ratings (including all item scores and a total score) in our database was 43,328. Mean (±SD) HoNOS total was 11.13 (±6.31, min 0, max 43). This large database was used to establish reliable item weights for Model 3 (see later).

2.2. Procedure for estimating missing item scores

To examine the accuracy of the three models, we used the aforementioned test set of 500 records, which comprised fully completed item scores (1 through 12) and their total. The mean age of the test set sample was 39.1 (±14.7) years and the ratio of male: female was 223:277 (45% male), which clearly reflected key demographic characteristics of the much larger sample. Minimum and maximum total HoNOS scores in the test set were, respectively, 0 and 34 (mean 11.36).

Using this test set, we systematically removed each of the 12 HoNOS items, one at a time by deleting all 500 item-scores in the respective spreadsheet column, and then used three different approaches, all outlined below, to estimate the missing item value for each patient.

2.3. Model 1: modal value of remaining items

In Model 1, the most frequently occurring score within the remaining 11 items for each patient was used as an estimate for the missing item value. So for example, if we had removed item 1, and the scores for items 2–12 for Patient 1 were 2, 0, 2, 1, 2, 2, 3, 2, 1, 1, 2, then we would estimate the value of Item 1 for Patient 1 to be 2, because this value occurs more frequently than any other value for this patient. In cases where there was a tie between two values (bi-modal distribution), we took the mean of these two scores.

2.4. Model 2: mean value of remaining items

In this model, we simply averaged the 11 remaining item scores and used the result to predict the missing value. So returning to the previous example for Patient 1 again, the mean of their 11 remaining items approximates to 1.64 and so this value would be our estimate for Item 1. In the interests of precision, we did not round up to the nearest integer, but did round to 2 decimal places where necessary.

2.5. Model 3: weighted mean value of remaining items

To calculate weighted means, we derived the average score for each item across the much larger database (i.e., the 43,328 HoNOS records described earlier, which excluded the 500 test set records), divided this by the average total HoNOS score across the entire database and then multiplied the result by 12. To give a worked example, the average score for Item 1 across all 43,328 ratings was 0.746 while the average total score across all ratings was 11.127. Thus, the weighting derived for Item 1 would be approximately 0.80 (i.e., [0.746/11.127] × 12). To then calculate the weighted *mean*, we would use the approach already outlined in *Model 2* but multiply the result by the respective weight, so the missing item value in the previous example would be estimated at $1.64 \times 0.8 = 1.32$. Again, in the interests of precision we did not round up estimated scores to the nearest integer, but worked to 2 decimal places. Table 2 below shows the weights derived for each of the 12 items using the above method. It is immediately apparent that the average response profile is far from uniform across all items: indeed there is considerable variation with some items tending to elicit scores that are

Table 2
Item weights generated from our large dataset of HoNOS scores. These weights are used to estimate the score of a given missing item by multiplying the respective weight by the mean of all those items for which a rated score does exist.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Weight	0.8	0.61	0.62	0.51	1	0.51	1.93	2.16	1.32	0.97	0.5	1.05

approximately 4 times greater than those on other items (for example, compare the weights for items 8 and 11).

2.6. Procedure for testing models

For all 500 patient records in our test set, we systematically removed one HoNOS item at a time and replaced it with the estimated value derived from each of the three models described above. So there were 500 such calculations for each HoNOS item, under each of the 3 models (500 × 3 × 12 in total). To examine model accuracy for each item, we compared the new estimated total derived for each patient under each model with the known (actual) total score. This is identical to the approach used in Gale and Hawley [13]. So for each of the 3 Models detailed above, there were 12 different statistical comparisons, each including 500 estimated vs. 500 actual total scores. We used pairwise *t*-tests to make these comparisons, with the expectation that the most accurate of the three models would yield the smallest number of statistical differences between average actual and average estimated totals.

3. Results

The average known total HoNOS score for our 500 patient test set was 11.36 points. Table 3 shows how each of the three models (Mode, Mean, Weighted Mean) performed when each of the 12 HoNOS items was systematically removed and then replaced with the estimates from each model. The closer the estimated average is to the known mean total of 11.36, the better the model is at predicting missing item scores.

In 10/12 statistical comparisons (each using data from 500 patients), Models 1 and 2 (Mode and Mean) generated estimates that were *significantly different* from the known totals. There is a strong tendency for Model 1 (Mode) to underestimate the value of missing items, and although Model 2 (Mean) does not do this, it is similarly inaccurate (the most inaccurate estimates were 1.78 points and 1.22 points respectively - and bear in mind that this is just the *average* discrepancy across 500 cases, not the maximum discrepancy). By contrast, Model 3 (Weighted Mean) produced much closer estimates to the known total score, deviating by a maximum of just 0.25 points (when item 8 was removed), and with the rest all being within 0.1 point. Consequently, Model 3 returned 11/12 non-significant comparisons and is by far the most accurate model.

Table 3
A comparison of the three models in predicting missing HoNOS item scores. The degrees of freedom value for all *t* values shown here is 499 (i.e. *n*-1). It is notable that the Weighed Mean model produces estimated total scores that are generally indiscernible from the known total score.

Item	*1. Mode*			*2. Mean*			*3. Weighted mean*		
	Est.	*t*	*p*	Est.	*t*	*p*	Est.	*t*	*p*
1	10.97	6.96	0.0001	11.62	5.9	0.0001	11.43	1.59	0.112
2	11.17	3.27	0.001	11.73	7.74	0.0001	11.35	0.21	0.838
3	11.13	3.81	0.0002	11.69	6.19	0.0001	11.32	0.77	0.443
4	11.29	1.21	0.23	11.86	11.55	0.0001	11.36	0.07	0.945
5	10.84	7.84	0.0001	11.42	1.09	0.28	11.42	1.09	0.28
6	11.18	2.98	0.003	11.77	8.19	0.0001	11.29	1.56	0.12
7	9.89	22.95	0.0001	10.46	17.14	0.0001	11.27	1.57	0.118
8	9.58	27.48	0.0001	10.14	22.37	0.0001	11.11	3.89	0.0001
9	10.41	16.22	0.0001	11.02	6.78	0.0001	11.31	0.867	0.387
10	10.74	11.29	0.0001	11.36	0.02	0.986	11.33	0.58	0.56
11	11.37	0.16	0.88	11.93	13.99	0.0001	11.43	1.88	0.06
12	10.66	12.83	0.0001	11.31	0.97	0.33	11.36	0.006	0.995

3.1. HoNOS for older people

Given the success of the Weighted Mean approach in predicting missing item data, we generated some additional weights for over 65s HoNOS data. In this age group, the HoNOS is also used regularly but elicits a very different profile of responding. Since we also had access to HoNOS data for a cohort of elderly patients from the same NHS Trust (*n* HoNOS 65+ records = 32,542), we calculated a new set of item weights for use with this age group. The weights are noticeably very different to those reported earlier (in Table 2) for the general adult cohort.

Given that the item weights in Table 4 show even greater diversity than those in Table 2 (the largest weight here is 27 times greater than the smallest, whereas for the general adult weights, this factor was 4.3), we can be very confident that the weighted mean approach would be the most accurate of the three models for this dataset too.

4. Discussion

In this brief paper, we have contrasted three simple statistical approaches to estimating missing HoNOS item values. In our experience it is quite common for HoNOS ratings to be incomplete and so we hope these results will have practical value for healthcare organisations who use this scale data routinely. The use of HONOS data for service improvement at a local level has already been demonstrated, although challenges of using routinely collected data have been identified [10,14].

Of the three models we compared, the Weighted Mean approach gave by far the most accurate results and this is because it successfully captures the average item-to-item variation seen in a typical HoNOS assessment. The same kind of model was also found to be a superior predictor for missing items in two commonly used depression scales [13]. Like these scales, the response profile for HoNOS tends to produce higher scores on some items than others, rather than a uniform pattern. We hope that the item weights and proposed calculations here will be of use to organisations dealing with large volumes of HoNOS data. Clinical teams may also find the weights useful if they need a rapid estimate for an individual patient assessment, although we note that most digital versions of the HoNOS require an integer score to be entered for each item so, at an individual patient level, the accuracy of this approach may be reduced by the requirement to round up or down to the nearest integer.

Of course, the three approaches tested in this study are not the only possible methods of estimating missing item scores. An alternative and relatively simple approach would be for the assessor could look at previous HoNOS assessments for a patient and use item scores from the most recent one to interpolate for any missing values in the current one, effectively a last observation carried forward approach. Whether this would generate more accurate results is untested but there are two problems with it. Firstly, it would not work if a previous rating did not exist and secondly, it may be rendered less accurate with the passage of time, especially if there were items that had repeatedly been left unrated. By contrast, the method proposed here can be applied using the data from just a single patient assessment, and can easily be utilized for large volumes of data by setting up a formula within a spreadsheet. We would reiterate here that the set of weights to be used is determined by the age group of the patient(s). For general adult caseloads, the weights in Table 2 should be used, while for over 65s, the weights in Table 4 will apply – and the respective weight must then be multiplied by the mean of all remaining rated items.

Unfortunately we did not have sufficient data available to generate missing item weights for HoNOS data within other patient groups (e.g. children or people with learning disabilities). It would, in principle, be possible to generate even more specific sets of weights, for example based on primary diagnosis. However the main aim of this study was to provide weights that can be applied to large volumes of data, which will inevitably include a broad case mix. As already noted, more complex algorithms for predicting missing data may be available via specialist software, but the advantage of the method proposed here is that it is free and can be easily implemented by anyone. With this in mind, we acknowledge that the case-mix and range of HoNOS scores may exhibit regional variations, and that the geographical area our data were collected from will have different patterns of deprivation to other counties (24.5% of Hertfordshire is more deprived than the national average, according to 2015 measures published by National Statistics). Although we believe that the item weights provided in this paper will be stable and widely applicable, there is no reason why other healthcare organisations should not derive their own weightings using the method outlined here and, indeed, it would be helpful to know how these compare.

We have not tested this model's accuracy when more than one item is missing but we know from the previous study with depression rating scales [13], that the Weighted Mean approach held up better than the two alternatives when more than one item was missing. And since the level of item-to-item variation is greater within HoNOS (compared to both the BDI and MÅDRS), a weighted means model will be the most accurate overall, irrespective of the number of missing items. Indeed, the weighted means approach could be applied to almost any psychiatric or medical rating scale, and is likely to be a more accurate approach for any scale that has a non-uniform pattern of response, than simply taking the mean or mode of residual items.

Item 8 (other mental and behavioural problems – to be specified) was the least well-estimated HoNOS item, being the only one where a significant difference emerged between actual and estimated scores using the weighted means model. From our database of 43,328 HoNOS ratings, it is notable that Item 8 had by far the highest mean score and standard deviation of all HoNOS items (2.01 ± 1.26), suggesting that it is much more likely than any other item to have scores falling across the full range of possible values, thereby rendering it less predictable. This may arise because of the special domain of item 8 that differs in how it is defined compared to the other rated items. The item 8 construct is in fact an overarching heading for a variety of possible other problems that an individual may experience. To rate, the clinician chooses a domain to be assessed (e.g., anxiety, sleep, obsessive-compulsive, sexual, and a number of other health problems), and these will vary considerably in both type and severity, from case to case. In reality then, the population mean for item 8 is the mean score across a range of differently rated, and in many ways unrelated, constructs. This is unlike the other 11 HoNOS items, which rate a single construct. Nonetheless, even for Item 8, the estimated mean differed from the actual mean by just a quarter of one point, so the Weighted Means model is still reasonably accurate for this item. Finally on this point, the test set comprised 500 cases so power to detect small differences (≥0.25 points per item) in our sample is relatively high (>0.85).

Of course there should be no substitute for collecting accurate ratings in the first place. A scale completed via a thorough assessment, and containing no missing data, should always be the gold standard. However, in cases where this is not possible, the method of estimation described here may be a useful second best, and may also help some organisations to make better use of the historical data they have already collected.

The HoNOS itself has come under some recent criticism, most notably for a lack of association between scores on 11/12 items and actual service costs (regular vs. high), and this has led to speculation about whether the HoNOS is an appropriate tool for implementing Payment by Results in England [10]. However, it is notable that in Twomey et al.'s study, missing item data in groups of more severely ill patients was a limitation, so replication of the lack of association between HoNOS items and service use is warranted. So, although there may be concerns about the validity of the scale for Payment by Results, it remains a ubiquitous current assessment and any means of improving rating accuracy ought to be of some value to healthcare organisations.

Table 4
Item weights generated from a large dataset of HoNOS scores (65+ years). Again, these weights are used to estimate the value of a missing item by multiplying the respective weight value by the mean of all items that have been previously rated.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Weight	0.73	0.08	0.09	2.78	2.18	0.45	0.99	1.12	0.40	2.03	0.24	0.90

Finally, it is important to remember that HoNOS has utility other than for Payment by Results, and was originally developed as a mental health outcome tool. Defining a set of outcome measures for mental health is notoriously difficult given the highly variable nature of symptoms and associated problems. HoNOS is one recognised method of understanding and reporting on outcomes and, because of the drive to use it to inform Payment by Results in the English NHS, has been very widely used across the country. However research making use of this data set is sparse, and it is possible that more could be understood about the nature of mental illness and mental health services in the U.K. if this data was examined more closely. Aside from testing statistical models for missing item prediction, this study demonstrates that it is possible to access large sets of HoNOS data from NHS organisations and to ask questions of it. We encourage other researchers to make more use of this relatively untapped resource for the benefit of people with mental health problems.

Declaration of Competing Interest

None.

Acknowledgements

The authors are grateful to Hertfordshire Partnership University NHS Trust, and specifically Avi Reddy for facilitating access to the anonymous HoNOS data. We are also grateful for the helpful comments of two anonymous reviewers.

References

[1] Royal College of Psychiatrists (RCPsych). Health of the Nation Outcome Scale. https://www.rcpsych.ac.uk/events/in-house-training/health-of-nation-outcome-scales.

[2] 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.

[3] Self R, Rigby A, Leggett C, Paxton R. Clinical decision support tool: a rational needs-based approach to making clinical decisions. J Ment Health 2008;17(1):33–48.

[4] Clark M. Mental health care clusters and payment by results: considerations for social inclusion and recovery. LSE Research Online http://eprints.lse.ac.uk/37529/1/Mental_health_care_clusters_and_payment_by_results(lsero).pdf; 2011.

[5] NHS England and NHS Improvement. https://improvement.nhs.uk/documents/599/Annex_C_-_Mental_Health_Clustering_bookletl.pdf; 2016.

[6] James M, Painter J, Buckingham B, Stewart MW. A review and update of the Health of the Nation Outcome Scales (HoNOS). Br J Psychiatr Bull 2018;42(2):63–8.

[7] NCCMH. https://www.rcpsych.ac.uk/docs/default-source/improving-care/nccmh/nccmh-eip-guidance.pdf?sfvrsn=ae517db4_2; 2016.

[8] NCCMH. Dementia. https://www.rcpsych.ac.uk/docs/default-source/improving-care/nccmh/nccmh-dementia-care-pathway-full-implementation-guidance.pdf?sfvrsn=cdef189d_6; 2018.

[9] NHS England and NHS Improvement. https://www.england.nhs.uk/mentalhealth/wp-content/uploads/sites/29/2016/02/mh-quality-outcome.pdf; 2016.

[10] Twomey C, Prina AM, Baldwin DS, Das-Munshi J, Kingdon D, Koeser L, et al. Utility of the Health of the Nation Outcome Scales (HoNOS) in predicting mental health service costs for patients with common mental health problems: historical cohort study. PLoS ONE 2016. https://doi.org/10.1371/journal.pone.0167103.

[11] Montgomery SA, Åsberg MA. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979;134:382–9.

[12] Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:53–63.

[13] Gale TM, Hawley CJ. A model for handling missing items on two depression rating scales. Int Clin Psychopharmacol 2001;16(4):2005–14.

[14] Crawford MJ, Zoha M, MacDonald AJD, Kingdon D. Improving the quality of mental health services using patient outcome data: Making the most of HoNOS. Br J Psychiatr Bull 2017;41(3):172–6.