

BMJ Open Measurement properties of the Health of the Nation Outcome Scales (HoNOS) family of measures: protocol for a systematic review

Meredith G Harris,^{1,2} Claudia Sparti,^{1,2} Roman Scheurer,^{1,2} Tim Coombs,³ Jane Pirkis,⁴ Torleif Ruud,^{5,6} Steve Kisely,^{7,8,9,10} Ketil Hanssen-Bauer,^{5,6} Johan Siqueland,^{5,6,11} Philip M Burgess^{1,2}

To cite: Harris MG, Sparti C, Scheurer R, *et al*. Measurement properties of the Health of the Nation Outcome Scales (HoNOS) family of measures: protocol for a systematic review. *BMJ Open* 2018;**8**:e021177. doi:10.1136/bmjopen-2017-021177

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2017-021177>).

Received 17 December 2017
Revised 8 February 2018
Accepted 19 February 2018

ABSTRACT

Introduction The Health of the Nation Outcome Scales (HoNOS) for adults, and equivalent measures for children and adolescents and older people, are widely used in clinical practice and research contexts to measure mental health and functional outcomes. Additional HoNOS measures have been developed for special populations and applications. Stakeholders require synthesised information about the measurement properties of these measures to assess whether they are fit for use with intended service settings and populations and to establish performance benchmarks. This planned systematic review will critically appraise evidence on the measurement properties of the HoNOS family of measures.

Methods and analysis Journal articles meeting inclusion criteria will be identified via a search of seven electronic databases: MEDLINE via EBSCOhost, PsycINFO via APA PsycNET, Embase via Elsevier, Cumulative Index to Nursing and Allied Health Literature via EBSCOhost, Web of Science via Thomson Reuters, Google Scholar and the Cochrane Library. Variants of 'Health of the Nation Outcome Scales' or 'HoNOS' will be searched as text words. No restrictions will be placed on setting or language of publication. Reference lists of relevant studies and reviews will be scanned for additional eligible studies. Appraisal of reliability, validity, responsiveness and interpretability will be guided by the CONSensus-based Standards for the selection of health Measurement INstruments checklist. Feasibility/utility will be appraised using definitions and criteria derived from previous reviews. For reliability studies, we will also apply the Guidelines for Reporting Reliability and Agreement Studies to assess quality of reporting. Results will be synthesised narratively, separately for each measure, and by subgroup (eg, treatment setting, rater profession/experience or training) where possible. Meta-analyses will be undertaken where data are adequate.

Ethics and dissemination Ethics approval is not required as no primary data will be collected. Outcomes will be disseminated to stakeholders via reports, journal articles and presentations at meetings and conferences.

PROSPERO registration number CRD42017057871.

INTRODUCTION

In 1992, UK's *Health of the Nation* strategy set a target to 'improve significantly the health and social functioning of mentally ill

Strengths and limitations of this study

- This systematic review will apply structured checklists that standardise the appraisal of available evidence.
- The review potentially extends previous systematic reviews on the topic by including meta-analyses of relevant measurement property metrics, if data are adequate.
- The review focuses on clinician-rated versions of the Health of the Nation Outcome Scales (HoNOS) family of measures; it does not include self- or proxy-completed versions.
- To reduce potential language bias, the review will include all relevant studies regardless of language of publication and studies using translated versions of the HoNOS family of measures.
- The search strategy does not include dissertations and reports; this may mean that a small amount of relevant information is missed.

people'.¹ The Health of the Nation Outcome Scales (HoNOS)² was developed under the auspices of the Royal College of Psychiatrists as a means of quantifying progress against this target. The HoNOS was developed as a clinician-rated measure for use with working-age adults in contact with mental health services. It comprises 12 scales assessing behaviour, impairment, symptoms and social functioning. Each scale is scored on a five-point scale representing maximum severity over the rating period, typically the previous 2 weeks (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). Scoring is guided by a glossary that provides specific anchor points for each scale. Subscale and total scores can be derived. Scores are based on clinical judgement, and can be used to guide patient treatment, resulting in the



For numbered affiliations see end of article.

Correspondence to

A/Prof. Meredith G Harris;
meredith.harris@uq.edu.au

provision of targeted care and support for clinical decision-making.³ In addition to its application in clinical practice, the HoNOS has also been used by researchers and policy-makers to monitor mental health service quality and effectiveness.⁴

In the decade following the development of the HoNOS, it was acknowledged that additional variants were required for use with specific populations.^{5 6} This resulted in the development of the HoNOS for Children and Adolescents (HoNOSCA),⁵ and the HoNOS for older adults (HoNOS 65+).⁶ Since then, the HoNOS family of measures has been extended to include the following: the HoNOS for adults confined to a secure facility (HoNOS-secure)⁷; the HoNOS for Acquired Brain Injury (HoNOS-ABI)⁸ and the HoNOS for People with Learning Disabilities (HoNOS-LD).⁹ Other variants of the HoNOS have been developed for administrative and research purposes, including the HoNOS for Payment by Results (HoNOS PbR)¹⁰ for casemix classification. All HoNOS measures apply the same scale scoring approach, but the number and content of the scales, and the subscale and total score structures, are tailored to the population or purpose (see online supplementary appendix 1 for an overview of the HoNOS measures).

HoNOS measures are widely used in clinical practice and research contexts. The HoNOS, HoNOSCA and HoNOS 65+ are now the most widely used routine outcome measures in mental health services in England.^{4 11} In Australia, these three measures have been implemented in inpatient, residential and ambulatory settings within public sector mental health services¹² and the HoNOS is used in private hospitals with psychiatric beds.¹³ In New Zealand, the HoNOS, HoNOSCA, HoNOS 65+, HoNOS-LD and HoNOS-secure have been mandated for routine collection in mental health services.¹⁴ Elsewhere, the routine implementation of HoNOS measures has occurred in local contexts, or is under active consideration. For example in Canada, the HoNOS, HoNOSCA and HoNOS 65+ are used in at least two provinces.¹⁵ In the Netherlands HoNOS, HoNOSCA and HoNOS 65+ have been used by various mental health services and are among instruments recommended by benchmarking systems for routine use.¹⁶ In Germany, HoNOS has been used in the sector of rehabilitative mental healthcare.¹⁷ In Norway, the national Norwegian Patient Register is preparing for use of HoNOS and HoNOSCA as possible routine outcome measures,¹⁸ and the HoNOSCA is used routinely within several child and adolescent mental health services in Denmark, Sweden and Norway. Various HoNOS measures have been translated into languages including Norwegian, Danish, Dutch, Spanish, Italian, Greek, German, Lithuanian, French and Thai. The HoNOS measures were designed as clinician-rated measures, although self- and proxy-completed versions of the HoNOS, HoNOSCA and HoNOS 65+ have also been developed.^{19–21}

Implementing routine outcomes monitoring in mental health services involves a substantial commitment of

resources in training and education, data management, and analysis and reporting.¹² It is therefore important that the measurement properties (ie, their reliability, validity and responsiveness or sensitivity to change) and practical aspects (ie, their interpretability and feasibility/utility in practice) of selected measures are acceptable.^{22 23}

Consumers, carers, clinicians, managers, policy-makers and researchers require up-to-date, synthesised evidence about the performance of the HoNOS measures to help them decide whether these measures are fit for use with the intended service settings and populations. Systematic reviews can inform such decision-making, as they provide an opportunity to compare findings from individual studies on a 'level playing field' and consider the reasons for agreements and disagreements. They may provide an opportunity to establish benchmarks for measurement property metrics, against which clinicians can compare the measures' performance in their own environment. Researchers also require this information to support their choice of measures when reporting or designing new studies.²⁴

A preparatory scoping search located a number of reviews^{25–32} that have sought to systematically identify and evaluate available information on the measurement properties of one or more members of the HoNOS family of measures. These reviews report that, for the most part, the HoNOS measures demonstrate acceptable performance on various measurement properties, and a number have identified HoNOS measures as suitable candidates for routine outcomes monitoring for certain populations and purposes when evaluated against a set of predetermined criteria.^{25–27 33} These reviews have also been useful in highlighting areas for potential improvement. For example, reviews have shown that the four HoNOS scales measuring aspects of social functioning (relationships, activities of daily living, living conditions, and occupation and activities) tend to perform less well than the other individual scales and other subscales and the total score.²⁵ Reviews have highlighted areas of conflicting evidence—for example, with respect to the factor structure of the HoNOS,²⁸ HoNOSCA²⁸ and HoNOS 65+³¹ and aspects of their feasibility or utility.^{28 31 34} Reviews have also highlighted the absence of data, at particular points in time, for certain measurement properties—for example, the content validity of the HoNOSCA,²⁸ test–retest reliability and construct validity of the HoNOS 65+,^{28 31} the test–retest reliability of the HoNOS-secure^{26 29} and the internal consistency, test–retest reliability, and criterion validity of the HoNOS-LD.³²

There is now more than 20 years of accumulated knowledge about the performance of the HoNOS and its variants, but there is no single, up-to-date evaluation of a comprehensive range of measurement properties for all HoNOS measures. Given the widespread implementation and policy importance of the HoNOS measures, and their ongoing development, this would be an important reference point for all stakeholders. The most comprehensive review of the HoNOS,

HoNOSCA and HoNOS 65+ was conducted by Pirkis and colleagues²⁸ who examined evidence regarding their reliability, validity, responsiveness, interpretability and feasibility/utility from studies published up to 2005, regardless of treatment setting or population. A later review supplemented these results with studies published up to 2011, and expanded coverage to include the HoNOS-LD and HoNOS-secure, but did not consider evidence of interpretability and feasibility/utility for any measure.³² Other subsequent reviews have been undertaken pertaining to the HoNOS^{25 27 30 33} and the HoNOSCA,²⁷ but have been more narrow in scope, focused on: studies from a single population subgroup^{27 31 33}; a subset of HoNOS scales²⁵; or on one or two measurement properties only.^{27 30 33} The HoNOS-secure has been the subject of two additional reviews, one considering literature up to 2011²⁹ and the other up to 2015.²⁶ However, being a relatively new measure, the number of available studies was small.

The existing reviews of the measurement properties of the HoNOS family of measures are, collectively, also limited in a number of methodological aspects. First, existing reviews have often excluded studies published in languages other than English or been unclear about whether they have been excluded. Given the number of languages that HoNOS measures have been translated into, excluding papers published in languages other than English may significantly under-represent the body of available evidence. Second, none of the previous reviews have applied detailed structured checklists^{35 36} that standardise the appraisal of the quality of evidence on measurement properties. The benefits of structured checklists are that they provide guidance regarding which measurement properties are important and how to investigate them, and increase the likelihood that extracted results will be comparable across raters and studies.³⁶ Third, all of the previous reviews have been descriptive; none have sought to apply meta-analytic techniques to pool information across studies on measurement property metrics. This has been done with other measures.^{24 37–39} Doing so may improve confidence in the reliability of findings, if they are replicated across multiple studies, and may increase the external validity of findings, if findings are maintained when samples of varying composition are pooled.^{40 41} Fourth, the existing reviews have not formally examined whether measurement properties differ between subgroups defined by patient or study characteristics; factors such as patient mix and rater profession/experience or training may be important sources of variation.

The current systematic review will be undertaken to address the aforementioned gaps. Its objectives are to:

- determine the extent of available evidence regarding the measurement properties of clinician-rated measures in the HoNOS family, in relation to their reliability, validity, responsiveness (or sensitivity to change), interpretability and feasibility/utility (or acceptability);
- examine the measurement properties of scales, subscales and totals for each measure in the HoNOS family, and whether performance varies across subgroups (eg, defined by treatment setting, clinical grouping, age group and rater profession/experience or training);
- apply meta-analytic techniques to generate pooled estimates of relevant measurement property metrics (eg, intraclass correlation coefficient (ICC), kappa, Cronbach's alpha and Pearson's correlation coefficient) for each clinician-rated HoNOS measure's scales, subscales and totals (by subgroup), where possible.

METHODS AND ANALYSIS

Design and registration

The review protocol was lodged with the PROSPERO (International Prospective Register of Systematic Reviews) on 22 February 2017 (CRD42017057871) and updated on 23 November 2017. It was developed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols (PRISMA-P) statement⁴² (see online supplementary appendix 2). The completed review will be reported according to the PRISMA guidelines.⁴³ Protocol amendments will be documented on PROSPERO with the date of all amendments, and a description of the changes and the rationale.

Search strategy

We will identify relevant peer-reviewed journal articles via a search of seven electronic databases from their inception: MEDLINE via EBSCOhost, PsycINFO via APA PsycNET, Embase via Elsevier, Cumulative Index to Nursing and Allied Health Literature via EBSCOhost, Web of Science via Thomson Reuters, Google Scholar and the Cochrane Library. Variants of the following terms will be searched as text words or keywords (as appropriate): 'Health of the Nation Outcome Scales' or 'HoNOS' (see online supplementary appendix 3 for the full search strategy). Searches were initially undertaken in January 2017 and were updated in February 2018. Exploratory searches of non-English language databases (China Academic Journals of China National Knowledge Infrastructure, Scientific Electronic Library Online via Web of Science, Russian Citation Index via Web of Science and Korean Citation Index—Korean Journal Database via Web of Science) were also undertaken, but these did not improve capture of eligible studies so were not incorporated into the search strategy. Reference lists of relevant articles will be scanned for additional studies. We will also contact the Royal College of Psychiatrists' Research Unit in the UK for details of any additional publications they may be aware of.

The search will focus on identifying original published studies in the form of articles published in peer-reviewed journals. It is possible that some relevant studies may be published in doctoral theses and government or

other reports; however, these sources are more difficult to search for comprehensively. Exploratory searches for thesis and report material yielded only a small amount of material that was not also published in peer-reviewed journal articles, hence we excluded these sources from the search strategy. Database search results, including abstracts and full-text articles, as relevant, will be downloaded into the EndNote reference management software package (EndNote V.X5.0.1).

Selection of studies

Eligibility criteria

A study will be included if all the following six criteria are fulfilled: (1) it is published in full in a peer-reviewed journal (ie, abstracts, letters and other short communications will be excluded unless they provide sufficient information to make the majority of required ratings); (2) its primary aim is to assess relevant measurement properties (within the domains of reliability, validity and responsiveness, as well as interpretability or feasibility/utility) for any of the HoNOS family of measures including: HoNOS, HoNOSCA, HoNOS 65+, HoNOS-secure, HoNOS-LD, HoNOS-ABI, HoNOS PbR or another HoNOS measure whose development has been described in the literature; (3) the HoNOS measure is clinician rated, regardless of mode of completion (eg, paper-and-pencil, directly entered into electronic medical record); (4) the clinical characteristics of participants are consistent with the HoNOS measures' target population; (5) relevant results are not duplicated elsewhere (where the same data are reported in more than one source, the more comprehensive version will be included) and (6) a full text version can be obtained. We will restrict studies to those with the primary aim of developing or testing the measurement properties of a HoNOS measure, as evidence from studies designed for other purposes may be difficult to interpret.⁴⁴ There will be no restrictions placed on study design, setting, context of study or language of publication. Studies using any language versions of HoNOS measures will be included.

Screening process

Two reviewers will independently screen the titles and abstracts yielded by the search against the eligibility criteria. Full-text articles will be obtained when one of the two reviewers concludes that the abstract/titles indicate that the HoNOS was potentially used in the study or when abstracts are not available. Relevant excerpts from articles published in languages other than English will be translated into English. Two reviewers will independently read the full-text articles and decide whether they meet the eligibility criteria. Disagreements will be resolved via discussion or, where consensus cannot be reached, with recourse to a third review author.

Appraisal of the methodological quality of included studies

The primary outcomes of this review will be the measurement properties of the HoNOS measures. The scope

of measurement properties considered in this review is based on the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) taxonomy.³⁶ In this taxonomy, measurement properties are clustered into domains—reliability (internal consistency, test–retest reliability, inter-rater reliability and measurement error), validity (content validity, construct validity and criterion validity) and responsiveness. We will include interpretability, which is not considered as a measurement property in COSMIN, but is an important characteristic of a measurement instrument. We will also include feasibility/utility which is important when assessing items for use in routine service delivery contexts. The definitions for each domain, measurement property and specific aspects to be assessed are shown in [table 1](#).

The appraisal of the methodological quality of the evaluation of measurement properties in included studies will be undertaken using the COSMIN checklist.^{36 45 46} The checklist includes 12 boxes; each box corresponds to a measurement property or aspect and contains between 4 and 18 items; each item is scored on a four-point rating scale (excellent, good, fair, poor) guided by descriptive anchor points.⁴⁷ The items focus on study attributes specific to the evaluation of measurement properties, including items that assess the adequacy of sample sizes, and whether appropriate statistical tests were performed. An overall quality score for a given measurement property is determined by taking the lowest rating assigned to any constituent item. For measurement properties where there is a particular item that is consistently under-reported, we may consider reporting the overall quality ratings with and without the inclusion of the item, informed by existing precedents (eg, Sitnikova *et al*⁴⁸). A quality score will not be derived for feasibility/utility because it is not included in COSMIN. For studies that used item response theory (IRT) methods, an additional set of items is provided for recording whether the requirements for IRT models have been met. These ratings are then taken into account when assigning an overall rating score on each measurement property assessed in that study.

Complementary information about the quality of reporting of reliability studies will be evaluated using the Guidelines for Reporting Reliability and Agreement Studies (GRRAS)⁴⁹ GRRAS comprises 15 guidelines that should be followed when agreement and reliability are reported. These cover the reporting of title and abstract (one guideline), introduction (four guidelines), methods (five guidelines), results (three guidelines), discussion (one guideline) and auxiliary material (one guideline). GRRAS guidance is narrative only; the guidelines are not accompanied by standardised system for coding the information extracted from included studies.

Data extraction

Data will be extracted from each study into a series of prespecified templates. For each study, we will capture information about:

Table 1 Definitions of domains, measurement properties and aspects of measurement properties

Domain	Measurement property (and aspect)	Definition
Reliability		'The degree to which the measurement is free from measurement error.' ³⁶
	Internal consistency	'The degree of the interrelatedness among the items.' ³⁶
	Test–retest reliability	'The proportion of the total variance [in repeated measurements of patients who have not changed] which is because of true differences among patients.' ³⁶
	Inter-rater reliability	'The proportion of total variance [in repeated measurements by different persons on the same occasion] which is because of true differences among patients.' ³⁶
	Measurement error	'The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured.' ³⁶
Validity		'The degree to which [the] instrument measures the construct(s) it purports to measure.' ³⁶
	Content validity	'The degree to which the content of [the instrument's scales and subscales] is an adequate reflection of the construct to be measured.' ³⁶
	Face validity	'The degree to which (the items of) [the] instrument indeed looks as though they are an adequate reflection of the construct to be measured.' ³⁶
	Construct validity	'The degree to which the scores of [the] instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other measures or differences between relevant groups) based on the assumption that the instrument validly measures the constructs to be measured.' ³⁶
	Structural validity	'The degree to which the scores of (the) instrument are an adequate reflection of the dimensionality of the construct to be measured.' ³⁶
	Hypothesis testing (includes convergent, discriminant and known groups validity)	As per construct validity. ³⁶
	Cross-cultural validity	'The degree to which the performance of the items on a translated or culturally adapted [...] instrument are an adequate reflection of the performance of the items of the original version of the [...] instrument.' ³⁶
	Criterion validity (concurrent, predictive or postdictive)	'The degree to which the scores on [the] instrument are an adequate reflection of a "gold standard".' ³⁶
Responsiveness		'The ability of [the] instrument to detect change over time in the construct to be measured.' ³⁶
Interpretability*		'The degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations—to [the] instrument's quantitative scores or change in scores.' ³⁶
	Percentage of missing items	
	Description of how missing items were handled	
	Floor and ceiling effects	'Percentage of respondents with the lowest possible score (floor effects) and highest possible score (ceiling effects).'
	Average scores	'Reported baseline values for [the instrument's scales and [or] subscales].' ²⁴
Minimally important change	'Smallest change in a score that is considered to be important.'	
Feasibility/utility*		'Degree to which the instrument is acceptable to and useful for stakeholders.' ²⁸
	Time to complete	'Reported time taken to complete [the instrument].' ²⁴

Continued

Table 1 Continued

Domain	Measurement property (and aspect)	Definition
	Completion rate	Percentage of clinicians completing the instrument in practice. ²⁸
	Stakeholder perceptions	May include views regarding: usefulness in treatment planning and clinical decision-making; usefulness in service evaluation and benchmarking; attitudes to routine outcome measurement; organisational/infrastructure barriers/enablers; training and feedback mechanisms; and protocols for assessment. ^{22 27 28}

*Interpretability and feasibility/utility are not considered measurement properties but are important characteristics of a measurement instrument.

- ▶ publication information (including authors, year);
- ▶ questionnaire characteristics (including the HoNOS measure(s) under evaluation; version number if relevant; items used in the study or modifications made to the measure);
- ▶ study characteristics (year(s) of study entry; country; language of measure(s); study design; setting; eligibility criteria; assessment occasion(s); sampling method; treatment/intervention type; experimental or routine conditions; and mode of HoNOS measure completion).

Some studies report multiple findings for a given measurement property, based on different subgroups (defined by patient characteristics, setting, assessment occasion, or other study characteristics) or using more than one metric. Therefore for each subgroup, we will capture information about:

- ▶ subgroup characteristics (including clinical profile; mean age and SD or age range; percentage female; treatment setting; point in care; and rater profession/experience and training in rating HoNOS);
- ▶ details of analysis (including number of patients involved; whether results pertain to scale, subscale or total score; metric reported and analyses undertaken);
- ▶ results reported (values on an appropriate metric or a narrative statement of results, as appropriate).

The COSMIN checklist with four-point rating scale will be completed for each study. For studies reporting on reliability, additional templates will be developed to capture the information required to assess the study against the GRRAS. For example, to assess the quality of reporting against GRRAS guideline 13 'report estimates of reliability and agreement including statistical measures of uncertainty', information to be extracted will include: the reliability or agreement metric used; whether or not uncertainty was reported; whether a sample size calculation was performed; whether results were pooled across subgroups and, if so, the pooling procedure and evidence of heterogeneity.

Data extraction will be undertaken independently by two reviewers; discrepancies that cannot be resolved via discussion will be referred to a third review author. Where necessary, we will contact authors for clarification of published data. The data extraction templates will be

piloted on several studies and modified as needed before use.

Methodological appraisal

Methodological appraisal of the measurement properties within the reliability, validity and responsiveness domains will be interpreted using the adequacy criteria proposed by the COSMIN group,⁵⁰ modified as necessary drawing on precedents set by previous reviews^{24 51–53} For each measurement property examined in a given study, the reported findings will be compared with a statistical threshold for adequacy and rated on a three-point system indicating whether or not it meets the threshold ('+', positive rating; '?', indeterminate rating; '-', negative rating). By way of example, the original COSMIN adequacy criteria for inter-rater reliability are: '+' ICC or weighted kappa ≥ 0.70 ; '?' ICC or weighted kappa not reported; and '-' ICC or weighted kappa < 0.70 . For interpretability and feasibility/utility, information regarding relevant aspects will be evaluated against available thresholds or, where thresholds are not available, via a narrative synthesis^{24 50}

Data synthesis

We will start by conducting a narrative synthesis of the information captured in the templates, comparing and contrasting methodological parameters and results across studies, and over time. Following COSMIN recommendations,⁴⁴ a best evidence synthesis will then be undertaken at the measurement property level. First, an overall rating (positive, indeterminate, negative) of the adequacy of each measurement property will be made. Second, a level of evidence (strong, moderate, limited, conflicting, unknown) of the methodological quality of the included studies for each measurement property will be assigned. The ratings will be guided by the original criteria developed by Terwee and colleagues⁴⁷ and any relevant modifications.²⁴ Qualitative findings regarding interpretability and feasibility/utility will be synthesised narratively. For reliability studies, quality of reporting information will be synthesised narratively, guided by the GRRAS.⁴⁹

In undertaking these syntheses, we will initially consider information from different language versions of the same measure separately, as it cannot be assumed that different

language versions will have the same measurement properties.⁵³

In addition, where the volume of available data is adequate, meta-analyses will be undertaken using Comprehensive Meta-Analysis software V.3 to generate pooled estimates of relevant measurement property metrics (eg, ICC, kappa, Cronbach's alpha and Pearson's correlation coefficient) for individual scales, subscales and total scores. Syntheses will involve combining data at the level reported in the paper (eg, individual scale, subscale, 10-item total or 12-item total), not across levels. It is not our intention to synthesise data across variants of the HoNOS (eg, HoNOSCA and HoNOS-LD) or across versions of a given HoNOS measure (eg, if modifications have been made).

We will start by including all relevant values on the properties of interest in a meta-analysis and then investigate whether heterogeneity is lower for subgroups that would be expected to be more homogenous by design. We will test for statistical heterogeneity using the I^2 statistic and, where present, will use the random effects model. If indicated, outcomes for subgroups of interest (eg, treatment setting, rater profession/experience or training) will be explored using meta-regression or subgroup analysis. Where there are more than 10 studies for a particular outcome we will test for publication bias using funnel plots. To examine the impact of study quality, we will conduct sensitivity analyses by excluding studies of 'poor' quality according to the COSMIN checklist. Where pooling of data across studies is not feasible, we will undertake narrative syntheses to explore differences in findings across subgroups, and according to study quality.

ETHICS AND DISSEMINATION

No primary data will be collected, therefore ethics approval is not required. We intend to summarise the outcomes of this study in one or more reports and/or presentations to the funder of this review. It is also planned that outcomes will be reported in peer-reviewed journal publications and in presentations at academic conferences and other meetings.

Author affiliations

¹School of Public Health, The University of Queensland, Brisbane, Australia

²Policy and Epidemiology Group, Queensland Centre for Mental Health Research, Brisbane, Australia

³Illawarra Institute for Mental Health, University of Wollongong, Wollongong, Australia

⁴Centre for Mental Health, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia

⁵Division of Mental Health Services, Akershus University Hospital, Lørenskog, Norway

⁶Institute of Clinical Medicine, University of Oslo, Oslo, Norway

⁷School of Medicine, The University of Queensland, Brisbane, Australia

⁸Metro South Addiction and Mental Health Service, Brisbane, Australia

⁹Department of Psychiatry, Dalhousie University, Halifax, Canada

¹⁰Department of Community Health and Epidemiology, Dalhousie University, Halifax, Canada

¹¹Division for Health Services at the Norwegian Institute of Public Health, Oslo, Norway

Contributors MGH and PMB are the guarantors and they conceived the study. MGH, PMB and JP designed the study, and CS, RS, TC, TR, SK, KH-B and JS provided critical inputs. The search strategy was developed by MGH, PMB, JP, CS, RS and SK. The selection criteria were developed by MGH, PMB, JP and CS. The analysis strategy was developed by MGH, PMB, CS, SK, KH-B and JS. Preliminary searches were conducted by RS and CS. Data extraction templates were developed by MGH and CS. MGH and CS drafted the protocol manuscript, and RS, TC, JP, TR, SK, KH-B, JS and PMB provided critical revisions. The final draft of the protocol manuscript was approved by all authors.

Funding This work is being led by the Australian Mental Health Outcomes and Classification Network which is funded by the Australian Government Department of Health.

Disclaimer The funding source had no involvement in the study design; collection, analysis and interpretation of data; the writing of the protocol; or in the decision to submit the protocol for publication.

Competing interests MGH, CS, RS, TC, JP, SK and PMB report that they or their institution received funding from the Australian Government Department of Health to conduct this work. TR, KH-B and JS have nothing to disclose.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2018. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

REFERENCES

1. Department of Health. *The Health of the Nation: A Strategy for Health in England*. London: Her Majesty's Stationery Office, 1992.
2. Wing JK, Beevor AS, Curtis RH, et al. Health of the Nation Outcome Scales (HoNOS). Research and development. *Br J Psychiatry* 1998;172:11–18.
3. Prowse L, Coombs T. The use of the Health of the Nation Outcome Scales (HoNOS) to inform discharge and transfer decisions in community mental health services. *Aust Health Rev* 2009;33:13–18.
4. Crawford MJ, Zoha M, Macdonald AJD, et al. Improving the quality of mental health services using patient outcome data: making the most of HoNOS. *BJPsych Bull* 2017;41:172–6.
5. Gowers SG, Harrington RC, Whitton A, et al. Brief scale for measuring the outcomes of emotional and behavioural disorders in children. Health of the Nation Outcome Scales for children and Adolescents (HoNOSCA). *Br J Psychiatry* 1999;174:413–6.
6. Burns A, Beevor A, Lelliott P, et al. Health of the nation outcome scales for elderly people (HoNOS 65+). *Br J Psychiatry* 1999;174:424–7.
7. Dickens G, Sugarman P, Walker L. HoNOS-secure: A reliable outcome measure for users of secure and forensic mental health services. *J Forens Psychiatry Psychol* 2007;18:507–14.
8. Coetzer R, Du Toit PL. HoNOS-ABI: a clinically useful outcome measure? *Psychiatr Bull* 2001;25:421–2.
9. Roy A, Matthews H, Clifford P, et al. Health of the Nation Outcome Scales for People with Learning Disabilities (HoNOS-LD). *Br J Psychiatry* 2002;180:61–6.
10. Department of Health. *The Mental Health Clustering Booklet (V3.0) (2013/14)*. 2013. <https://www.gov.uk/government/publications/mental-health-payment-by-results-arrangements-for-2013-14> (accessed 23 Aug 2017).
11. *Royal College of Psychiatrists. What is HoNOS?* London, 2017. <http://www.rcpsych.ac.uk/trainingsychiatry/conferencetraining/resources/honos/whatishonos.aspx>. (accessed 22 Aug 2017).
12. Burgess P, Pirkis J, Coombs T. Routine outcome measurement in Australia. *Int Rev Psychiatry* 2015;27:264–75.
13. Morris-Yates A, Page AC. Routine measurement of outcomes by Australian private hospital-based psychiatric services. Trauer T, ed. *Outcome Measurement in Mental Health: Theory and Practice*. United Kingdom: Cambridge University Press, 2011:149–63.
14. Smith M, Baxendine S. Outcome measurement in New Zealand. *Int Rev Psychiatry* 2015;27:276–85.

15. Kisely S, Adair CE, Lin E, *et al.* Routine outcome measures in Canada. *Int Rev Psychiatry* 2015;27:286–95.
16. Delespaul PA. Routine outcome measurement in the Netherlands - A focus on benchmarking. *Int Rev Psychiatry* 2015;27:320–8.
17. Puschner B, Becker T, Bauer S. Routine outcome measures in Germany. *Int Rev Psychiatry* 2015;27:329–37.
18. Ruud T. Routine outcome measures in Norway: Only partly implemented. *Int Rev Psychiatry* 2015;27:338–44.
19. Stewart M. Service user and significant other versions of the Health of the Nation Outcome Scales. *Australas Psychiatry* 2009;17:156–63.
20. Ausin B, Muñoz M, Pérez-Santos E. Development of the self-administered Spanish version of the health of the nation outcome scales for older adults: HoNOS65+A. *Eur Psychiatry* 2010;25:164–71.
21. Gowers S, Levine W, Bailey-Rogers S, *et al.* Use of a routine, self-report outcome measure (HoNOSCA-SR) in two adolescent mental health services. *Br J Psychiatry* 2002;180:266–9.
22. Delaffon V, Anwar Z, Noushad F, *et al.* Use of Health of the Nation Outcome Scales in psychiatry. *Advances in Psychiatric Treatment* 2012;18:173–9.
23. Mokkink LB, Prinsen CA, Bouter LM, *et al.* The COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther* 2016;20:105–13.
24. Collins NJ, Prinsen CA, Christensen R, *et al.* Knee Injury and Osteoarthritis Outcome Score (KOOS): systematic review and meta-analysis of measurement properties. *Osteoarthritis Cartilage* 2016;24:1317–29.
25. Burgess PM, Harris MG, Coombs T, *et al.* A systematic review of clinician-rated instruments to assess adults' levels of functioning in specialised public sector mental health services. *Aust N Z J Psychiatry* 2017;51:338–54.
26. Keulen-de Vos M, Schepers K. Needs assessment in forensic patients: a review of instrument suites. *Int J Forensic Ment Health* 2016;15:283–300.
27. Kwan B, Rickwood DJ. A systematic review of mental health outcome measures for young people aged 12 to 25 years. *BMC Psychiatry* 2015;15:279.
28. Pirkis JE, Burgess PM, Kirk PK, *et al.* A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures. *Health Qual Life Outcomes* 2005;3:76.
29. Shinkfield G, Ogloff J. A Review and Analysis of Routine Outcome Measures for Forensic Mental Health Services. *Int J Forensic Ment Health* 2014;13:252–71.
30. Smith W, Patel A, McCrone P, *et al.* Reducing outcome measures in mental health: a systematic review of the methods. *J Ment Health* 2016;25:461–72.
31. Turner S. Are the Health of the Nation Outcome Scales (HoNOS) useful for measuring outcomes in older people's mental health services? *Aging Ment Health* 2004;8:387–96.
32. Te Pou. *The HoNOS Family of Measures: A technical review of their psychometric properties.* Auckland: Te Pou o Te Whakaaro Nui, 2012.
33. Demarchi C, Bohanna I, Baune BT, *et al.* Detecting psychotic symptoms in Indigenous populations: a review of available assessment tools. *Schizophr Res* 2012;139:136–43.
34. Jacobs R. *Investigating Patient Outcome Measures in Mental Health.* CHE Research Paper 48. York: Centre for Health Economics, University of York, 2009.
35. Aaronson N, Alonso J, Burnam A, *et al.* Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–205.
36. Mokkink LB, Terwee CB, Patrick DL, *et al.* The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
37. Chiarotto A, Maxwell LJ, Terwee CB, *et al.* Roland-Morris Disability Questionnaire and Oswestry Disability Index: Which Has Better Measurement Properties for Measuring Physical Functioning in Nonspecific Low Back Pain? Systematic Review and Meta-Analysis. *Phys Ther* 2016;96:1620–37.
38. Broadbent E, Wilkes C, Koschwanec H, *et al.* A systematic review and meta-analysis of the Brief Illness Perception Questionnaire. *Psychol Health* 2015;30:1361–85.
39. Scaini S, Battaglia M, Beidel DC, *et al.* A meta-analysis of the cross-cultural psychometric properties of the Social Phobia and Anxiety Inventory for Children (SPAI-C). *J Anxiety Disord* 2012;26:182–8.
40. Gliner JA, Morgan GA, Harmon RJ. Meta-analysis: formulation and interpretation. *J Am Acad Child Adolesc Psychiatry* 2003;42:1376–9.
41. Hale WW, Crocetti E, Raaijmakers QA, *et al.* A meta-analysis of the cross-cultural psychometric properties of the Screen for Child Anxiety Related Emotional Disorders (SCARED). *J Child Psychol Psychiatr* 2011;52:80–90.
42. Shamseer L, Moher D, Clarke M, *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015;349:g7647.
43. Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
44. Terwee CB. *Protocol for systematic reviews of measurement properties.* Amsterdam: VU University Medical Center, 2011. <http://www.cosmin.nl/images/upload/files/Protocol%20klinimetriche%20review%20version%20nov%202011.pdf>. (accessed 23 Aug 2017).
45. Mokkink LB, Terwee CB, Knol DL, *et al.* The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010;10:22.
46. Mokkink LB, Terwee CB, Patrick DL, *et al.* The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010;19:539–49.
47. Terwee CB, Mokkink LB, Knol DL, *et al.* Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;21:651–7.
48. Sitnikova K, Dijkstra-Kersten SMA, Mokkink LB, *et al.* Systematic review of measurement properties of questionnaires measuring somatization in primary care patients. *J Psychosom Res* 2017;103:42–62.
49. Kottner J, Audigé L, Brorson S, *et al.* Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
50. Terwee CB, Bot SD, de Boer MR, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.
51. Schellingerhout JM, Heymans MW, Verhagen AP, *et al.* Measurement properties of translated versions of neck-specific questionnaires: a systematic review. *BMC Med Res Methodol* 2011;11:87.
52. Uijen AA, Heinst CW, Schellevis FG, *et al.* Measurement properties of questionnaires measuring continuity of care: a systematic review. *PLoS One* 2012;7:e42256.
53. Heini D, Prinsen CA, Drucker AM, *et al.* Measurement properties of quality of life measurement instruments for infants, children and adolescents with eczema: protocol for a systematic review. *Syst Rev* 2016;5:25.