Intern Emerg Med (2007) 2:315–317
DOI 10.1007/s11739-007-0086-y

COCHRANE'S CORNER

Scales to climb borderline personalities: when science goes nowhere

L. Moja • G. Virgili • A. Liberati
G.F. Gensini • R. Gusinu • A.A. Conti

L. Moja (✉) • A. Liberati
Italian Cochrane Centre
Mario Negri Institute for Pharmacological Research
Via Eritrea 62
I-20157 Milan, Italy
e-mail: moja@marionegri.it

G. Virgili
Clinica Oculistica
Azienda Ospedaliero-Universitaria Careggi
Florence, Italy

G.F. Gensini
Department of Critical Care Medicine and Surgery
University of Florence and Azienda Ospedaliero-Universitaria
Careggi
Florence, Italy

R. Gusinu
DAI Cardiologico e dei Vasi, Azienda Ospedaliero-Universitaria
Careggi
Florence, Italy

A.A. Conti
Department of Critical Care Medicine and Surgery
University of Florence and Don Carlo Gnocchi Foundation
IRCCS Florence, Florence, Italy

Published online: 17 December 2007

The methodologist's point of view

L. Moja, A. Liberati, G. Virgili

Measures of disease and measurement tool

The Cochrane review "Psychological therapies for people with borderline personality disorder (BPD)" by Binks et al. [1] highlights a peculiarity of outcome measures in psychology, which is also found in social sciences: the widespread use of several types of scales which aim at targeting different aspects of mental health, subjective well-being or the general burden of illness.

Although these psychosocial scales are increasingly used, scepticism and confusion remain regarding the ability of many scales to summarise composite indices and if they are meaningful in orienting a conclusive clinical decision. We explore some of the limitations regarding scales use and misuse in science.

The purpose of the review by Binks et al. [1] is to evaluate the evidence for the efficacy of psychological intervention for people with BPD. This review is extremely important from a policy standpoint given the prevalence of the problem (2%), the association with deliberate self-harm and suicide, and the long-term cost to the healthcare system due to chronicity and intense health service demands by people with BPD.

The Authors expected that outcomes would not have been consistently reported across the included trials and that different measures of disease and scales had been used (e.g., quality of life, mental state, behaviour, etc.). In order not to miss useful observations, the Authors included data from any rating scale as long as these instruments had been described in a peer-reviewed journal. This broad approach resulted in 15 outcome categories, each of which has several measures, totalling 82 outcomes. There are six primary outcomes, placed in five different categories. For example, the category *mental state* is articulated in: (1) *general mental state*; (2) *not clinically important change in general mental state* (a primary outcome); (3) *not any change in general mental state*; and … etc. Death is the only outcome with no subcategories.

The advantage of the choice made by the Authors is that the whole body of research on this topic is covered, but at what price? The disadvantage is that a general reader will not go through this review as he will find himself lost trying to understand the essence of the trial results. Even a psychologist using a methodological approach will be puzzled by the question: *Which is an effective treatment for borderline personality disorders*? This is not easy to answer given the multitude and fragmentation of the instruments and outcomes presented in the review.

Multiple comparisons

After an extensive search the Authors selected seven relevant randomised control trials, each enrolling between 23 and 64 patients, for a total of 262 people with an average of 37 people per study. The small size of trials relates to a general lack of statistical power of the studies. This means that it is difficult to show differences between the intervention and the control groups. To avoid the risk of insignificant findings, trialists used a large number of rating scales to measure outcomes. Furthermore the majority of the trials analysed the scale and subscales at many different time intervals (<6 months, by 6–12 months, by 18 months), exponentially increasing the number of comparisons. Performing multiple comparisons easily generates type I errors, achieving statistical significance on specific measures with small numbers [2, 3]. It is likely that all positive findings have been reported in the BPD literature, whilst negative results have often been omitted, leading to reporting bias. The meta-analyses in this review suffer from the type of this bias, as reported by Binks et al.

Clinical relevance

The clinical relevance of some of the rating tools adopted in the trials is questionable. For example, to evaluate mental state at least four scales resulting in complex index have been used: the Beck Depression Inventory, the Spielberger Anger Expression Scale, the Symptoms Check List and the Beck Anxiety Inventory. Although all scales have been previously cited in peer-review medical journals, we do not have basic information about their validation. These instruments combine information from numerous questionnaire items that span from well defined behaviour or objective functioning ('In the last 6 months did you attempt to kill yourself?') to subjective health appraisal ('Do you feel your family cannot understand you?'). Furthermore, these scales variously combine frequency, severity and/or duration of symptoms with no pre-specified priorities among these attributes. Some focus on one psychological attribute, such as depression or anger, while others collapse the above attributes into one estimate. We are aware that the complexity of human personality can never be satisfactorily expressed as one score on any scale. We can, however, make meaningful estimations of some human attributes, but we can do that only for one attribute at time (often referred to as unidimensionality) [4]. Confounding a number of attributes into a single generic score makes confident predictions from that score more hazardous and the score an inconsistent summary of the disease. On the basis of a complex score, it is difficult to clinically predict if an individual (or a trial arm) is worsened or ameliorated, as the diagnosis relies more on an illogical process and chance than on anything objective and scientific. Not knowing whether different attributes are equally correlated may increase the risk of data dredging and distorted reporting in any attempt to demonstrate post-hoc differences between interventions [3]. This problem, again, is magnified by the limited size of most included trials.

Scale validity and reliability

Questionnaires exploring medical disease, which allow decisions to be made about a healthcare intervention, must be designed with respect to three key issues: reliability, validity and referring population [5]. Reliability is the extent to which the measurements of a test remain consistent over repeated testing of the same subject under identical conditions. A scale is reliable if it yields consistent results of the same measure. It is unreliable if repeated measurements yield different results. A high level of reliability is particularly important when the effect of an intervention on psychological distress is measured using a pre-test/post-test design. In this type of research, design and pre-test and post-test reliability are fundamental to the credibility of results and the ability to attribute differences in pre-test and post-test performance to the intervention being tested. The ability to attribute such changes is also affected by research design.

In psychology the validity of a research instrument is the degree to which the instrument measures what it is supposed to measure. Validity is closely related to reliability because for an instrument to be valid, it must be reliable. It is also important to remember that instruments may in fact be reliable even when they are not valid.

Finally it is important to know in which population the questionnaire has been validated. Most scales may vary with personality dimensions [5]. Prevalence of neuroticism or extraversion, for example, may lead to different response predispositions in various diagnostic groups. Indeed it is important that a scale has been validated in the diagnostic group in which you are interested.

In systematic reviews a decision about which scale to include in the final analyses may be procrastinated after the data have been collected, with a preference for standardised and validated scales.

How can we deal with complexity?

We agree with Binks et al. [1], who suggest the use of other clinically relevant outcomes such as hospital admissions, medication use or days off work in trials. These outcomes are generally more interpretable and are strongly related to patients' quality of life. We need more studies considering these meaningful outcomes and less inconsistent scales. A large treatment effect will be detected regardless of the instrument used to compare the quality of life. The problems arise when generic instruments are used in a broad range of patients with different diseases: these will be less responsive in the detection of treatment-induced changes.

Another option is to use a statistical approach such as the bivariate meta-analysis and its extensions. This technique simultaneously combines information on multiple outcome measures [6]. We can model each trial arm separately, fitting a model that assumes the study data to be bivariate normally distributed to investigate the relationship between true and surrogate outcomes. When the correlations among outcome variables are known and the included studies report true and surrogate outcomes, this can lead to increase in efficiency of estimation.

The Campbell Collaboration, the equivalent of the Cochrane Collaboration in the social science and education field, suggests that meta-analysts 'should not ignore the dependence among study outcomes'; however, they also note that 'the consequences of accounting for (modelling) dependence or ignoring it are not well understood' [7].

The predetermined choice of a few, relevant, patient-centred outcome measures remains the optimal way to try to deal with complexity.

Clinician's point of view

G.F. Gensini, R. Gusinu, A.A. Conti

Dealing with biological and clinical complexity is a daily challenge for health operators. In order to effectively tackle this problem, rating tools are increasingly elaborated and adopted in the field of healthcare. However, building a robust and reliable scale ensuring methodological correctness and clinical relevance is one of the most difficult undertakings for healthcare professionals.

In their interesting Cochrane review, Binks et al. [1] have considered and analysed the structured adoption of different scales aimed at highlighting various features not only of mental health, but also of subjective wellbeing, and, more generally, of the overall burden of disease. Their work deserves credit in many respects, also evidencing how several rating tools available today in the area of the evaluation of borderline personalities have a clinical relevance that is "improvable".

Ideally, assessment scales should be basically validated, should have acceptable sensitivity to changes in the severity of symptoms, appropriate inter-rater and test–retest reliability, and high internal consistency [4]. These methodological aspects are also decisive in defining the clinical applicability of the rating tools examined.

Unfortunately, as correctly underlined by Moja et al. in this article, even data on the real validation of the discussed rating scales are not available and consequently analysable. In the light of this basic drawback, and also considering that, to effectively propose reliable estimations of human dimensions, probably only one dimension at a time should be considered, from a clinician's point of view the careful pre-definition and identification of a limited number of really important clinical end points should be seen as an appropriate way of trying to tackle complexity in a meaningful manner. Clinicians, even those who are not directly involved in the mental health area, should be encouraged to actively participate in large registries to test different scales, so as to optimally combine the identifiable elements most useful in detecting patients at higher risk.

References

1. Binks CA, Fenton M, McCarthy L et al (2006) Psychological therapies for people with borderline personality disorder. Cochrane Database Syst Rev (1):CD005652. DOI: 10.1002/14651858.CD005652
2. Thornley B, Adams C (1998) Content and quality of 2000 controlled trials in schizophrenia over 50 years. BMJ 317:1181–1184
3. Pocock SJ (1997) Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. Control Clin Trials 18:530–545
4. Bond TG, Fox CM (2007) Applying the Rasch model: fundamental measurement in the human sciences. Lawrence Erlbaum Associates, Mahwah, NJ
5. Calvert MJ, Freemantle N (2004) Use of health-related quality of life in prescribing research. Part 2: Methodological considerations for the assessment of health-related quality of life in clinical trials. J Clin Pharm Ther 29:85–94
6. Sutton AJ, Higgins JPT (2007) Recent developments in meta-analysis. Stat Med (in press) DOI: 10.1002/sim.2934
7. Becker BJ, Hedges LV, Pigott TD (2004) Campbell Collaboration Statistical Analysis Policy Brief. A Campbell Collaboration resource document. http://www.campbellcollaboration.org/ECG/policy_statasp