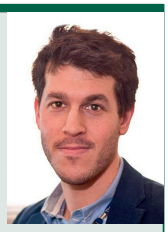
BJPsych Open (2015)
1, e5–e7. doi: 10.1192/bjpo.bp.115.000786

Editorial

High heterogeneity and low reliability in the diagnosis of major depression will impair the development of new drugs



Samuel M. Lieblich, David J. Castle, Christos Pantelis, Malcolm Hopwood, Allan Hunter Young and Ian P. Everall

Summary
Major depressive disorder is a common diagnosis associated with a high burden of disease that has proven to be highly heterogeneous and unreliable. Treatments currently available demonstrate limited efficacy and effectiveness. New drug development is urgently required but is likely to be hindered by diagnostic limitations.

Declarations of interest
D.J.C. has received grants and personal fees from Eli Lilly, Janssen-Cilag, Roche, Allergen, Bristol-Myers Squibb, Pfizer, Lundbeck, AstraZeneca, Hospira, Organon, Sanofi-Aventis, and Wyeth during the writing of this review. C.P. has received grant support from Janssen-Cilag, Eli Lilly, Hospira (Mayne), AstraZeneca, and received honoraria for consultancy to Janssen-Cilag, Eli Lilly, Hospira (Mayne), AstraZeneca, Pfizer, Schering Plough, and Lundbeck. Over the past 2 years he has participated on advisory boards for Janssen-Cilag and Lundbeck, and received honoraria for talks presented at educational meetings organised by AstraZeneca, Janssen-Cilag and Lundbeck. M.H. has received personal fees or grants from Lundbeck, AstraZeneca and Servier during the writing of this review. A.H.Y. reports personal fees from Lundbeck, Sunovion, AstraZeneca and Janssen outside the submitted work. I.P.E. has received personal fees or grants from Lundbeck, AstraZeneca, and Abbvie during the writing of this review.

Copyright and usage
© The Royal College of Psychiatrists 2015. This is an open access article distributed under the terms of the Creative Commons Non-Commercial, No Derivatives (CC BY-NC-ND) licence.

Dr Samuel M. Lieblich (pictured) is a registrar in psychiatry working at the Royal Melbourne Hospital. He is conducting research into the role of the von economo neuron in the pathogenesis of schizophrenia. Professor David J. Castle is Chair of Psychiatry at St Vincent's Health and Melbourne University. His broad clinical and research interests encompass schizophrenia and related disorders, bipolar disorder, cannabis misuse, OCD spectrum disorders and disorders of body image. Christos Pantelis is an NHMRC Senior Principal Research Fellow, Foundation Professor of Neuropsychiatry and Scientific Director of the Melbourne Neuropsychiatry Centre at Melbourne University and Melbourne Health. His work has focused on brain structural and functional changes during the transition to psychosis and other neurodevelopmental disorders. Malcolm Hopwood is the Ramsay Health Care Professor of Psychiatry, University of Melbourne. He specialises in clinical aspects of mood and anxiety disorders, psychopharmacology and psychiatric aspects of acquired brain injury. Professor Allan Young holds the Chair of Mood Disorders at the Institute of Psychiatry, Psychology and Neuroscience, King's College London where he is Director of the Centre for Affective Disorders and an Honorary Consultant in the Bethlem and Maudsley hospitals. Professor Ian Everall is the Cato Chair of Psychiatry and Head of the Department of Psychiatry, Melbourne University. He leads the Department in exploring molecular and cellular changes in the brain in major psychiatric disorders.

Major depressive disorder is a diagnostic category with apparently high prevalence, and evidently high heterogeneity, which is associated with a high burden of disease. In Europe, the major depressive disorder label is associated with the loss of 8% of all disability-adjusted life-years (DALYs) of 291 measured diseases.[1]

Nevertheless, diagnosis and treatment for people with this diagnosis are inadequate. Psychiatrists have a hard time agreeing on who does and does not have major depressive disorder: the field trials for DSM-5 demonstrated an intraclass kappa of 0.28.[2] This represents a value of 'minimal agreement', and means that highly trained specialist psychiatrists under study conditions were only able to agree that a patient has depression between 4 and 15% of the time.[2,3] Once patients are diagnosed, few receive evidence-based care[4] and those who do have a relatively low chance of recovery due to treatment.[5]

Clearly, new treatments – including new drugs – are required, but for new drug development to be successful, a less heterogeneous and more reliable clinical construct is required. Researchers must know what is meant by the label major depressive disorder and whether or not their trial participants have major depressive disorder to know whether their new drug is effective in its treatment. This is true for non-drug treatments also, although their relationship to diagnosis is distinct, and will not be addressed here. The major depressive disorder diagnosis, as it stands, is likely to include various clinical syndromes and various putative pathologies; it may therefore need to be sub-classified into more homogenous entities.

In 2013, the DSM-5 was released. It was intended to be a major overhaul of the DSM approach to the classification of mental illnesses. The American Psychiatric Association (APA) intended to use the massive output of neuroscientific research to produce psychiatric diagnoses that were valid. This did not happen and the diagnostic category of major depressive disorder appears to have become even less reliable. As these diagnostic criteria will be used in many upcoming drug trials to determine case and response conditions, the decreased reliability – which will manifest as more heterogeneous, less severe study cohorts – will be reflected in a reduction of the apparent efficacy of trial drugs. This reduction in efficacy may mirror the efficacy-to-effectiveness decrement seen in the passage of previously developed antidepressants from clinical trials to use in primary care. The result

https://doi.org/10.1192/bjpo.bp.115.000786 Published online by Cambridge University Press

could be a block in the drug development pipeline if no apparently efficacious drugs can be produced.

DSM unreliability

Diagnostic reliability is more important in the diagnosis of mental illnesses, when compared with other kinds of illness, because no gold standard exists (like trans-oesophageal echocardiography in mitral regurgitation) to confirm the validity of diagnosis. In 1974, Robert Spitzer, who led the production of the DSM-III, performed a meta-analysis of the extant reliability research[6] and found mean kappa values for common diagnoses. The various types of depression, involutional, neurotic and psychotic, had relatively low reliability with kappa values of 0.30, 0.26 and 0.24 respectively. Notwithstanding the potential for diagnostic disagreement between the types of depression, the reliability with which psychiatrists of the day determined the presence of any affective disorder whatsoever was also very low with a kappa of 0.41.

The field-testing of the DSM-III yielded higher kappa values. The kappa for major depression had improved considerably to 0.62[7] and 0.64 in DSM-III-R.[8] None of the major diagnostic domains became ostensibly less reliable with the advent of the DSM-III operationalised approach to diagnosis.[9]

No DSM-IV category was associated with markedly lower reliability than its equivalent DSM-III-R category, and there were some areas of modest improvement.[8] The diagnosis of major depression improved marginally but not significantly to a kappa value of 0.67–0.68.[8]

The DSM-5 field trials were carried out between 2010 and 2012 and consisted of 279 clinicians at 11 academic centres in the USA and Canada.[2] They showed decreased reliability in all major domains, with some diagnoses, such as mixed anxiety-depressive disorder (kappa=0−0.004), so unreliable as to appear useless in clinical practice. For major depressive disorder reliability was disconcertingly low with a kappa of 0.28.[2]

The apparent drop in reliability between DSM-IV and DSM-5 may be deceptive, as DSM-IV reliability trials were conducted on highly selected populations, using more controlled methods, than those used in the more pragmatic design of the DSM-5 field trials. Although the claim made by Regier *et al* in the field trial report,[2] that the DSM-IV and DSM-5 major depressive disorder diagnoses are identical, and that these results therefore reflect DSM-IV values under real-world conditions, is inaccurate: There are two major changes to the DSM-5 that have affected the reliability of the diagnosis of major depressive disorder,[2] and are likely to affect the observed efficacy of antidepressants in forthcoming drug trials: the removal of the clause 'The symptoms are not better accounted for by bereavement' from the major depressive disorder category and the creation of the chapter 'Depressive Disorders'. These are both significant changes to the way a clinician may arrive at a diagnosis of major depressive disorder, even if criterion A of the diagnostic criteria is unchanged. The creation of a new chapter for 'Depressive Disorders' separates major depressive disorder from bipolar affective disorder, reconfigures some DSM-IV diagnoses and introduces diagnostic categories that did not exist before. Dysthymic disorder has become persistent depressive disorder, and premenstrual dysphoria, which was previously only mentioned in the DSM-IV appendices, is now reified as premenstrual dysphoric disorder. These changes are likely to change the prevalence of the various disorders, and may therefore be one of the factors affecting the reliability observed in the field trials. The removal of the bereavement exclusion and the significant changes to overlapping diagnoses can be expected to have an effect on both diagnostic reliability and heterogeneity. Whether the DSM-5 created or perpetuated the unreliability problem, it cannot be used to determine the efficacy of antidepressant medication if it only helps researchers to distinguish between major depressive disorder and other illnesses 4 to 15% of the time.

Antidepressant effectiveness

The search for valid, or at least reliable, clinical diagnoses, has been undertaken in tandem with the search for effective pharmacological treatments for mental illnesses. Observed efficacy has been decreasing steadily in antidepressant trials over the past three decades.[5] A number of reasons have been proposed including: decreased placebo drop-out, increased study enrolment of less symptomatic patients recruited through advertisements and a more recent decrease in the publication bias that inflated earlier efficacy results.[5] Study design manipulation and other forms of researcher bias (like hypothesising after the result is known) probably continue to inflate efficacy results.[5]

It is important to recognise that trial efficacy is not the same as effectiveness in the 'real-world' of clinical practice. As primary care diagnosis has been more heterogeneous and less reliable than the strictly applied criteria of the DSM-IV, to see where antidepressant efficacy is headed, we can look to where real-world effectiveness has been in the DSM-IV era. The STAR*D study was a large US study aimed at determining antidepressant effectiveness in real-world settings, with highly heterogeneous real-world patients.[5] During this study, 4041 primary care patients who had sought treatment for depression were enrolled and treated with best-practice principles. During the 12-month follow-up period, only 1854 (45.9%) of patients experienced a remission of symptoms.

As antidepressant efficacy exceeds placebo efficacy only for a relatively high baseline severity of depression,[10] it may be that these heterogeneous samples include too many patients with mild symptoms. It follows that efficacy in future trials will drop further if the decreased reliability of the DSM-5 major depressive disorder diagnosis leads to the enrolment of less severe cases in antidepressant trials.

History and direction

The project of the APA expressed in the DSM-5 began in the late 1970s with the conception and publication of the DSM-III. Robert Spitzer, who along with Cohen and others developed the kappa statistic, led the production of that manual. His stated aim was to improve the validity (therefore utility) of mental health diagnosis by improving the reliability that imposed an upper limit on the validity of the system of classification. As he wrote in 1974: 'A necessary constraint on the validity of a system is its reliability'.[6] The reliability of psychiatric diagnoses was essential to the discipline to facilitate communication between clinicians, justify its place in the scientific medical community and to facilitate the prescription of a small number of drug classes, for a relatively large number of problems of living, behaving, thinking and feeling.

The DSM-III, DSM-III-R, DSM-IV and DSM-IV-TR did improve both the reliability and utility of diagnosis, and fostered increased confidence in the scientific aspirations (if not scientific basis) of psychiatry.

https://doi.org/10.1192/bjpo.bp.115.000786 Published online by Cambridge University Press

What was hoped for in the DSM-5 was the incorporation of a more scientific approach to symptom clustering that might yield a group of more homogeneous disorders that had hitherto been subsumed under the major depressive disorder label. Instead, the DSM-III era category of major depressive disorder was not only retained but also contentiously expanded by the removal of the clause excluding a diagnosis of major depressive disorder in a newly bereaved person. As is evident from the disappointing field trials, at best this change had no effect, at worst it is one of the causes of diminishing reliability. It is certainly plausible that it has significantly increased diagnostic heterogeneity: the removal of the bereavement exclusion means that in forthcoming trials a patient with anorexia, weight loss and severe psychomotor retardation who has recently attempted suicide by hanging can be enrolled in a drug trial alongside a patient whose spouse has died in the previous 2 weeks and complains of low mood, fatigue, insomnia and indecisiveness. The former patient fits better into the model of medical treatment and risk management encapsulated by contemporary psychiatry and can also be expected to have a better response to antidepressants.[11] The latter patient has previously been conceived of as enduring an unpleasant but normal, time-limited reaction to a significant event, and evidence that treatment with antidepressants is beneficial is critically impaired.[12]

If researchers engaged in the development of new drugs use the DSM-5 criteria to define caseness, they will be enrolling a mixed bag of patients with different clusters of symptoms, differing symptom severity, and different philosophical approaches to the designation of mental disorder.

Conclusions

That the DSM-5 was published at all despite field-testing results that demonstrated decreased reliability compared with previous editions does not support the APAs contention qua Spitzer, that the DSM project seeks validity through reliability. If that were so, reliability would have increased from DSM-III to DSM-5, or the field-test results would have been taken as evidence of failure. We contend that the approach to the field testing of the DSM-5, which has been more reflective of real-world practice, has both exposed the covert unreliability of the DSM-IV-TR diagnosis of major depressive disorder, and also exposed the fact that – at best – there has been no improvement in the DSM-5. That is, that major depressive disorder was unreliable then, and is more unreliable now. Ultimately, the heterogeneous construct currently known as major depressive disorder needs to be reconsidered. In the near term, where the DSM-5 is used in drug development, researchers must be aware of the possible changes to the composition of their study groups and address them in their study design, for instance, stratifying patients by severity; otherwise the altered major depressive disorder category may have the effect of bringing the published results for efficacy down to meet the poor primary care effectiveness results. If this happens, low published results for efficacy will result in a decreased yield of efficacious new drugs.

Samuel M. Lieblich, MBBS, MPsych, Department of Psychiatry, University of Melbourne, Melbourne, Australia; **David J. Castle**, MD, FRANZCP, St Vincent's Hospital Melbourne and The University of Melbourne and Faculty of Health Sciences, Australian Catholic University, Melbourne, Australia; **Christos Pantelis**, MD, MRCPsych, FRANZCP, Melbourne Neuropsychiatry Centre, The University of Melbourne and Melbourne Health and Florey Institute for Neuroscience & Mental Health, Melbourne, Australia, and Bedfordshire Centre For Mental Health Research in Association with the University of Cambridge, UK; **Malcolm Hopwood**, MD, FRANZCP, Professorial Psychiatry Unit, Albert Road Clinic, and University of Melbourne, Melbourne, Australia; **Allan Hunter Young**, Centre for Affective Disorders, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK; **Ian P. Everall**, BSc (Hons), MB ChB (Hons), PhD, DSc, FRCPsych, FRANZCP, FRCPath, Department of Psychiatry, University of Melbourne, Melbourne, Australia

Correspondence: Samuel M. Lieblich, Department of Psychiatry, University of Melbourne, Level: 01 Room: N10023, Main Block, Royal Melbourne Hospital, Parkville, Victoria 3052, Australia. Email: samuel.lieblich@unimelb.edu.au

First received 23 Apr 2015, final revision 8 Sep 2015, accepted 16 Oct 2015

Funding

C.P. was supported by an NHMRC Senior Principal Research Fellowship (628386).

References

1 Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012; **380**: 2197–223.

2 Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry* 2013; **170**: 59–70.

3 McHugh M. Interrater reliability: the kappa statistic. *Biochem Med* 2012; **22**: 276–82.

4 Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* 2003; **289**: 3095–105.

5 Pigott HE, Leventhal AM, Alter GS, Boren JJ. Efficacy and effectiveness of antidepressants: current status of research. *Psychother Psychosom* 2010; **79**: 267–79.

6 Spitzer RL, Fleiss JL. A re-analysis of the reliability of psychiatric diagnosis. *Br J Psychiatry* 1974; **125**: 341–7.

7 Williams JBW, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, et al. The structured clinical interview for DSM-III-R (SCID) reliability description of sites. *Arch Gen Psychiatry* 1992; **49**: 630–6.

8 Brown TA, Di Nardo PA, Lehman CL, Campbell LA. Reliability of DSM-IV anxiety and mood disorders: implications for the classification of emotional disorders. *J Abnorm Psychol* 2001; **110**: 49–58.

9 Spitzer RL, Forman JB, Nee J. DSM-III field trials: I. Initial interrater diagnostic reliability. *Am J Psychiatry* 1979; **136**: 815–7.

10 Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008; **5**: e45.

11 Carroll BJ. Bringing back melancholia. *Bipolar Disord* 2012; **14**: 1–5.

12 Corruble E, Falissard B, Gorwood P. Is DSM-IV bereavement exclusion for major depression relevant to treatment response? A case-control, prospective study. *J Clin Psychiatry* 2011; **72**: 898–902.

OPEN ACCESS

CC BY NC ND

https://doi.org/10.1192/bjpo.bp.115.000786 Published online by Cambridge University Press