The World Journal of Biological Psychiatry, 2008; 9(4): 242–247

EDITORIAL

How to grade categories of evidence

BORWIN BANDELOW[1], JOSEPH ZOHAR[2], SIEGFRIED KASPER[3] & HANS-JÜRGEN MÖLLER[4]

[1]*Department of Psychiatry and Psychotherapy, University of Göttingen, Göttingen, Germany,* [2]*Division of Psychiatry, Chaim-Sheba Medical Center, Tel-Hashomer, Ramat Gan, Israel,* [3]*Department of Psychiatry and Psychotherapy, Medical University of Vienna, Vienna, Austria, and* [4]*Department of Psychiatry and Psychotherapy, Ludwig Maximilian University, Munich, Germany*

Key words: *Guidelines, evidence, randomized controlled trials, clinical studies*

In this issue, the revised version of the World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for the Pharmacological Treatment of Anxiety, Obsessive-Compulsive and Post-Traumatic Stress Disorders are published (Bandelow et al. 2002). Although only 7 years have passed since the first publication of the guidelines, the Task Force now saw a need for a revised version, as new treatments have emerged and evidence for existing treatments has been consolidated. Also, all other guidelines of the WFSBP (Bauer et al. 2002a,b, 2007; Falkai et al. 2005, 2006; Grunze et al. 2002, 2003, 2004; Herpertz et al. 2007; Soyka et al. 2008) will be updated in future.

The present guideline is now based on over 500 evaluable controlled studies, and many more studies had to be evaluated for possible inclusion.

Clinicians are bombarded with "evidence" that a new treatment is more effective, has fewer side effects, faster onset of action or has other advantages. With an increasing number of treatment options available for patients with psychiatric disorders over the last decade and the growing body of evidence describing their efficacy and safety, clinicians often find it difficult to determine the best appropriate treatment for each patient. One aim of guidelines is to summarize and to simplify such findings, by carefully weighing advantages and disadvantages of the available treatment interventions. The main results are typically condensed into evidence categories, which are based on efficacy. However, not only efficacy is important for decision finding. For instance, of two treatments with the same efficacy, the one with the most benign side effect profile or the lowest costs should be recommended. The results of the guideline consensus process can be finally summarized in recommendation levels, which also take the risk–benefit ratio of the therapeutic interventions into account.

The various types of evidence can be arranged hierarchically in a grading system according to strength, with randomized controlled trials at the most definitive end and case reports or opinions of "respected authorities", which are not based on any published evidence ("eminence-based medicine"), at the least definitive end of the spectrum. When searching for a commonly used grading system of categories of evidence for the guidelines of the WFSBP, we found that there is no generally accepted system for medicinal or psychological treatment interventions. It would be desirable that the same hierarchy of evidence is used in all such guidelines. However, over 100 different systems exist for grading evidence, and none of these is preferred by most guideline panels. The reason for this diversity probably lies in the different requirements in different specialties in medicine. It may be difficult to construct a system that applies to all fields of medicine. In psychiatry, we are interested in whether a drug is better than placebo for treating depression or whether it can prevent relapses in bipolar disorder, whereas in surgery we are interested in whether a certain operation can prevent death from a rare cancer type; in internal medicine

Correspondence: Prof. Dr. med. Dipl.-Psych. Borwin Bandelow, Department of Psychiatry and Psychotherapy, University of Göttingen, von-Sieboldstr. 5, D-37075 Göttingen, Germany. E-mail: Sekretariat.Bandelow@med.uni-goettingen.de

ISSN 1562-2975 print/ISSN 1814-1412 online
DOI: 10.1080/15622970802456590

we may search for an optimal strategy to lower cholesterol levels to reduce the risk of cardiovascular disease. In all these different issues, different study types are necessary, and it is not easy to develop an evidence grading system for all these study types. There are efforts ongoing to find a general system for rating quality of evidence (Guyatt et al. 2008). However, for the WFSBP guidelines, we did not find a grading system that was adequate for the typical data in psychopharmacology. On the contrary, we found a number of problems with the existing systems, with the consequence that interventions with weak efficacy could be upgraded to the first level of evidence under certain circumstances.

For example, we found it difficult to adopt the system by Eccles and Mason (2001) which was used in the recent guidelines for panic disorder and generalized anxiety disorder of the U.K. National Institute for Clinical Excellence (NICE) (NICE 2007), due to some methodological issues. In Table I, this system is compared with the grading scheme of the WFSBP guidelines.

1. Level 1 evidence of the Eccles & Mason system requires "evidence obtained from a single randomized controlled trial or a meta-analysis of controlled trials". In this statement, it is not even required that this RCT has to be placebo controlled. This means that an underpowered RCT comparing a new drug with an established one and showing no difference could be seen as sufficient evidence, according to this phrasing. Moreover, even if a placebo controlled study exists, it is desirable not to base a first-level recommendation on one single study. The anticonvulsant valproate was effective in one very small double-blind placebo-controlled cross-over study in panic disorder – would this justify recommending valproate as first-line treatment for panic disorder? At least two studies, i.e. an independent replication of the initial study, should be a prerequisite for the best category of evidence.
2. Moreover, it would also be desirable that levels of evidence of guidelines are compatible with the requirements of the new drug approval authorities, e.g., the US Food and Drug Administration (FDA) or the European Medicines Agency (EMEA). The EMEA guidelines require three-arm trials, including a placebo arm and an active comparator. A recommended treatment intervention should not only be better than a pill or a psychological placebo, but also not less effective than an established treatment. However, in the Eccles & Mason guidelines, a drug that is inferior to a reference drug, but superior to placebo, would still reach Level 1. Therefore, it is evident that the first level of evidence should be reserved for treatments that are demonstrably at least of comparable efficacy as a reference treatment. However, this requirement is only applicable if such a standard treatment existed before.
3. The requirement of a meta-analysis for Level 1 is also problematic. Meta-analyses have some advantages. They may sometimes permit conclusions about efficacy to be drawn with a greater degree of confidence than is possible with qualitative reviews. When direct comparisons of two treatments are lacking, these can be compared by using meta-analysis, even when different rating scales have been used in these trials. When conflicting results exist for a certain treatment, meta-analysis can solve these discrepancies. While a single study is only powered for analysing the whole study population, a meta-analysis of many studies may have the statistical power to analyse smaller subgroups within the population (e.g., elderly patients). However, meta-analyses have a number of methodological shortcomings, which make them less reliable than the original studies:
   - Effect sizes are not easily comparable across different studies, when different efficacy measures are used. Even within a study, effect sizes may differ substantially, e.g., when comparing the results of the Hamilton Anxiety Scale with the Clinical Global Impression Scale for the same patients.
   - By combining many small studies to a large data set, the statistical power may increase to a sufficient magnitude to yield statistically significant results, but these effect sizes may be so small that they are meaningless for the patients. According to the Eccles & Mason system, a drug that was not superior to placebo in three well-powered studies could nevertheless reach Level 1 evidence after these studies were pooled in a meta-analysis, because of the artificially inflated power in the larger sample size.
   - In a meta-analysis, studies are included that differ substantially in patient selection, average illness severity, intervention, dosage, study duration, and outcome parameters. This may also be seen as an advantage, as the findings attain higher external validity. In real life, a new drug should work in all patients and not only in selected subgroups. However, generalizability of meta-analyses may also be a

Table I. Comparisons of the grading scheme for categories of evidence used by the NICE Guidelines for Anxiety Disorders (Eccles and Mason 2001) with the WFSBP Guidelines system. [1]These standards are defined in Bandelow et al. (2008, this issue).

Eccles and Mason (2001)		World Federation of Societies of Biological Psychiatry (WFSBP)	
Category of evidence	Description	Category of evidence	Description
I	Evidence from: – meta-analysis of randomised controlled trials, or – at least one randomised controlled trial	↑↑ A	Full Evidence From Controlled Studies is based on: two or more double-blind, parallel-group, randomized controlled studies (RCTs) showing superiority to placebo (or in the case of psychotherapy studies, superiority to a "psychological placebo" in a study with adequate blinding) *and* one or more positive RCT showing superiority to or equivalent efficacy compared with established comparator treatment in a three-arm study with placebo control or in a well-powered non-inferiority trial (only required if such a standard treatment exists) In the case of existing negative studies (studies showing non-superiority to placebo or inferiority to comparator treatment), these must be outweighed by at least two more positive studies or a meta-analysis of all available studies shows superiority to placebo and non-inferiority to an established comparator treatment. Studies must fulfill established methodological standards[1]. The decision is based on the primary efficacy measure.
		↑B	Limited Positive Evidence From Controlled Studies is based on: one or more RCTs showing superiority to placebo (or in the case of psychotherapy studies, superiority to a "psychological placebo") *or* a randomized controlled comparison with a standard treatment without placebo control with a sample size sufficient for a non-inferiority trial and no negative studies exist
II	Evidence from: – at least one controlled study without randomisation, or – at least one other type of quasi-experimental study	(↑) C	Evidence from Uncontrolled Studies or Case Reports/Expert Opinion
		C1	Uncontrolled Studies is based on: one or more positive naturalistic open studies (with a minimum of five evaluable patients) *or* a comparison with a reference drug with a sample size insufficient for a non-inferiority trial *and* no negative controlled studies exist
III	Evidence from non-experimental descriptive studies, such as comparative studies, correlation studies and case–control studies		
		C2	Case Reports is based on: one or more positive case reports *and* no negative controlled studies exist

Table I (*Continued*)

Eccles and Mason (2001)		World Federation of Societies of Biological Psychiatry (WFSBP)	
Category of evidence	Description	Category of evidence	Description
IV	Evidence from expert committee reports or opinions and/or clinical experience of respected authorities	C3	Based on the opinion of experts in the field or clinical experience
		↔ D	Inconsistent Results Positive RCTs are outweighed by an approximately equal number of negative studies
		↓ E	Negative Evidence The majority of RCTs studies shows non-superiority to placebo (or in the case of psychotherapy studies, superiority to a "psychological placebo") or inferiority to comparator treatment
		? F	Lack of Evidence Adequate studies proving efficacy or non-efficacy are lacking

problematic issue. For example, if a meta-analysis of studies with three SSRIs shows that these drugs are effective in treating an anxiety disorder, can these findings be generalized to all SSRIs, including those that have never been investigated in this special disorder – although all SSRIs are chemically different?

– In particular, when two different treatments have been investigated in different settings, results may be biased, as shown in this example: In a clinical trial at a university department of psychology, cognitive behavioural therapy (CBT) for panic disorder is compared with a wait list control. Participants having the luck to be selected randomly for CBT have a high positive expectancy that the treatment will improve their symptoms, because they are not blind to the treatment condition, CBT has a good reputation, and the therapists of the centre are well-known specialists in their field. Additionally, they are allowed to be kept on their previous medications, e.g., SSRIs and benzodiazepines, according to typical study protocols used in this kind of design. In a second study, panic patients take part in a double-blind trial with a new drug designed for licensing the drug. The participants have the expectancy that placebo is not effective and there is also a possibility that the new drug will not be effective. Both studies are methodologically sound. However, a meta-analysis comparing the effect sizes of both studies would be biased, as the effect sizes of the first study are inflated due to positive expectancy and additional drug effects, while expectancy in the second study is lowered.

– Meta-analyses are often contemptuously described as "garbage-in/garbage-out", meaning that excellent and flawed studies are mixed together in one analysis. Therefore, studies should only be selected when they fulfil certain methodological standards, regarding sample size, randomization, control group, dosage, rating scales, statistical methods, etc. However, by varying these methodological requirements, study selection may be biased, by including favoured studies and excluding other studies on the basis of putative flaws. This "cherry-picking" may be one of the most important reasons why meta-analyses of the same database often come to contradictory results. For example, different meta-analyses comparing CBT and drug therapy for panic disorder found either superiority of CBT over drug therapy or equal efficacy. Some found no gains from the combination of both, while others found a substantial advantage (Bandelow et al. 2007).

– The statistical power to detect differences between treatments is dependent on both the number of observations and the magnitude of the effect. This also applies to meta-analyses. In the case of conventional meta-analysis, N is the number of studies included. Thus the power of a meta-analysis of only two or three studies is limited, unless the effect sizes are large, which is unlikely in the case of studies in

anxiety disorders and other psychiatric illnesses.
  - Meta-analyses may in fact be unnecessary for the decision when a consistent database exists, e.g., when three or more trials unequivocally show a difference to placebo, which is mostly the case for drugs that are recommended as first-line drugs. Because of many shortcomings, meta-analyses should not be seen as the highest level of evidence (Maier and Möller 2005) and should only be used when a summary of the original studies is not sufficient to draw a definite conclusion.
  - An alternative for conventional meta-analyses are "pooled" analyses. In this kind of meta-analysis, original individual patient data from a number of studies are pooled, rather than calculating effect sizes from the mean, standard deviation, and sample size from a published paper. This requires access to the raw data, but today this is no major obstacle due to modern electronic transfer techniques. Pooled analyses provide highly reliable results, but may also have the problem of artificially inflated power.
4. While not differentiating sufficiently between ample evidence from a number of placebo-controlled and comparator trials and minimal evidence from just one RCT within Level 1, the Eccles & Mason categories have two categories for open studies. Level II includes "controlled studies without randomisation" and "quasi-experimental studies", while Level III includes "non-experimental descriptive studies". Studies without randomisation and double-blinding are outdated in psychopharmacology, due to placebo effects, other unspecific factors, and publication biases. When, for example, two groups of patients are compared retrospectively who have received two different drugs and the effect sizes are not significantly different, the scientific value of such a result is extremely limited. The same probably applies for "quasi-experimental studies", whatever is meant with this phrasing. Open studies may have some heuristic value and can stimulate further double-blind trials, but do not have a confirmative purpose. However, by differentiating between two kinds of poor quality studies, the first kind is upvalued without justification.
5. Comparisons with a reference drug with a sample size insufficient for a non-inferiority trial should not obtain a higher level than open studies (e.g., an RCT in which a drug was as effective as an established reference drug, but only 30 patients were included in each group). Although these studies are double-blind, efficacy of a drug cannot be concluded when only data from an underpowered comparison with a standard drug are available.
6. In the Eccles & Mason guidelines, a level of evidence is missing for treatments with inconsistent evidence, i.e. when controlled positive studies are outweighed by an approximately equal number of negative studies, for example, when a drug showed superiority in three studies but failed to do so in three other studies. This should become transparent to the healthcare provider. This is most probably due to a weak effect of the investigated drug. When all six studies are combined in a meta-analysis, the result may be a significant differentiation from placebo, and the drug would even reach Level 1 in the Eccles & Mason system, although the effect size is only marginal. Instead, this drug should fall into a category for "inconsistent results". Clinicians would not use the drug as first-line treatment, but in patients unresponsive to all standard treatments, this drug still could be an option.
7. In a treatment guideline, also interventions should be commented upon that are widely used in care primary, but were either shown to be ineffective in RCTs or were never investigated in this disorder. Absence of evidence is not the same as evidence of absence of an effect. For example, there is strong negative evidence against the use of beta-blockers in anxiety disorders, because all available studies showed non-superiority to placebo. However, when there is lack of evidence for one drug in a special indication, it could still be tried in desperate cases of patients unresponsive to standard treatments, while such an attempt should not be undertaken with a drug that definitely showed negative evidence.

Because of these shortcomings of existing grading systems, we decided to develop special levels of evidence for the WFSBP guidelines, by integrating suggestions from other guidelines and by trying to use definitions that are optimally adapted to the situation of evidential data in psychiatry, in order to provide optimal transparency for the users of this guideline. It is planned to use this grading system for categories of evidence for all future revisions of the guidelines of the World Federation of Societies of Biological Psychiatry (WFSBP).

Consensus guidelines may improve the overall quality of treatment (however, adherence to pub-

lished guidelines is not always satisfactory). Guidelines also may have some influence on the design of future studies. By insisting on high quality standards, they can stimulate the application of rigorous methodological standards.

References

Bandelow B, Zohar J, Hollander E, Kasper S, Möller HJ. 2002. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for the pharmacological treatment of anxiety, obsessive-compulsive and posttraumatic stress disorders. World J Biol Psychiatry 3:171–199.

Bandelow B, Seidler-Brandler U, Becker A, Wedekind D, Rüther E. 2007. Meta-analysis of randomized controlled comparisons of psychopharmacological and psychological treatments for anxiety disorders. World J Biol Psychiatry 8:175–187.

Bauer M, Whybrow PC, Angst J, Versiani M, Moller HJ. 2002a. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Unipolar Depressive Disorders, Part 2: Maintenance treatment of major depressive disorder and treatment of chronic depressive disorders and subthreshold depressions. World J Biol Psychiatry 3:69–86.

Bauer M, Whybrow PC, Angst J, Versiani M, Möller HJ. 2002b. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Unipolar Depressive Disorders, Part 1: Acute and continuation treatment of major depressive disorder. World J Biol Psychiatry 3: 5–43.

Bauer M, Bschor T, Pfennig A, Whybrow PC, Angst J, Versiani M, Möller HJ. 2007. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of unipolar depressive disorders in primary care. World J Biol Psychiatry 8:67–104.

Eccles M, Mason J. 2001. How to develop cost-conscious guidelines. Health Technol Assess 5:1–69.

Falkai P, Wobrock T, Lieberman J, Glenthoj B, Gattaz WF, Möller HJ. 2005. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of schizophrenia, Part 1: acute treatment of schizophrenia. World J Biol Psychiatry 6:132–191.

Falkai P, Wobrock T, Lieberman J, Glenthoj B, Gattaz WF, Möller HJ. 2006. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Schizophrenia. Part 2: Long-term treatment of schizophrenia. World J Biol Psychiatry 7:5–40.

Grunze H, Kasper S, Goodwin G, Bowden C, Baldwin D, Licht R, et al. 2002. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Bipolar Disorders. Part I: Treatment of bipolar depression. World J Biol Psychiatry 3:115–124.

Grunze H, Kasper S, Goodwin G, Bowden C, Baldwin D, Licht RW, et al. 2003. The World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for the Biological Treatment of Bipolar Disorders, Part II: Treatment of Mania. World J Biol Psychiatry 4:5–13.

Grunze H, Kasper S, Goodwin G, Bowden C, Möller HJ. 2004. The World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for the Biological Treatment of Bipolar Disorders. Part III: Maintenance treatment. World J Biol Psychiatry 5:120–135.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Br Med J 336:924–926.

Herpertz SC, Zanarini M, Schulz CS, Siever L, Lieb K, Moller HJ. 2007. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Personality Disorders. World J Biol Psychiatry 8:212–244.

Maier W, Möller HJ. 2005. Metaanalyses – highest level of empirical evidence? Eur Arch Psychiatry Clin Neurosci 255:369–370.

NICE. 2007. National Institute for Health and Clinical Excellence (NICE. Anxiety (amended): Management of Anxiety (Panic Disorder, with or without Agoraphobia, and Generalised Anxiety Disorder) in Adults in Primary, Secondary and Community Care.

Soyka M, Kranzler HR, Berglund M, Gorelick D, Hesselbrock V, Johnson BA, Moller HJ. 2008. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Substance Use and Related Disorders, Part 1: Alcoholism. World J Biol Psychiatry 9:6–23.