




WFSBP guidelines on how to grade treatment evidence for clinical guideline development

Alkomiet Hasan, Borwin Bandelow, Lakshmi N. Yatham, Michael Berk, Peter Falkai, Hans-Jürgen Möller, Siegfried Kasper & WFSBP Guideline Task Force Chairs


To cite this article: Alkomiet Hasan, Borwin Bandelow, Lakshmi N. Yatham, Michael Berk, Peter Falkai, Hans-Jürgen Möller, Siegfried Kasper & WFSBP Guideline Task Force Chairs (2019) WFSBP guidelines on how to grade treatment evidence for clinical guideline development, The World Journal of Biological Psychiatry, 20:1, 2-16, DOI: [10.1080/15622975.2018.1557346](https://doi.org/10.1080/15622975.2018.1557346)


To link to this article: <https://doi.org/10.1080/15622975.2018.1557346>

 View supplementary material [↗](#)


 Accepted author version posted online: 11 Dec 2018.
Published online: 04 Feb 2019.

 Submit your article to this journal [↗](#)

 Article views: 145



 View related articles [↗](#)

 View Crossmark data [↗](#)

 Citing articles: 1 View citing articles [↗](#)



WFSBP guidelines on how to grade treatment evidence for clinical guideline development

Alkomiet Hasan^{a*}, Borwin Bandelow^{b*} , Lakshmi N. Yatham^c, Michael Berk^{d,e} , Peter Falkai^a, Hans-Jürgen Möller^a, Siegfried Kasper^f and WFSBP Guideline Task Force Chairs[†]

^aDepartment of Psychiatry and Psychotherapy, Klinikum der Universität München, Ludwig-Maximilians Universität München, Munich, Germany; ^bDepartment of Psychiatry and Psychotherapy, Universitätsmedizin Göttingen, Goettingen, Germany; ^cVancouver Coastal Health and Providence Health Care, University of British Columbia, Vancouver, Canada; ^dIMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, Australia; ^eOrygen, The National Centre of Excellence in Youth Mental Health, the Florey Institute for Neuroscience and Mental Health, and the Department of Psychiatry, University of Melbourne, Parkville, Australia; ^fDepartment of Psychiatry and Psychotherapy, Medizinische Universität Wien, Vienna, Austria

ABSTRACT

Objective and methods: This paper reviews sources of data typically used in guideline development, available grading systems, their pros and cons, and the methods for evaluating risks of bias in publications, and proposes a revised method for grading evidence and recommendations for use in development of clinical treatment guidelines.

Results: The new World Federation of Societies of Biological Psychiatry (WFSBP) grading system allows guideline developers to follow a multi-step approach of defining levels of evidence, applying criteria for grading (define the acceptability) and the grading of recommendations.

Conclusions: Further, these updated WFSBP recommendations for rating evidence and treatment recommendations provide a grading system that takes into account potential biases in sources of evidence in arriving at final ratings that are likely more clinically meaningful and pragmatic and thus should be used for the development of future treatment guidelines.

ARTICLE HISTORY

Received 30 November 2018
Accepted 3 December 2018

KEYWORDS

evidence-based psychiatry; guidelines; grading evidence; meta-analysis; randomized controlled trial

1. Introduction

The volume of published clinical studies and meta-analyses in medicine in general and psychiatry in particular continues to increase at a rapid pace, and it is becoming increasingly difficult to keep abreast of the current evidence with regard to the various pharmacological and non-pharmacological treatment options. Many different methods of grading evidence in medicine have been proposed over the last decades, but there is no interdisciplinary consensus with regard to which of these methods is the best. In 2008, the World Federation of Societies of Biological Psychiatry (WFSBP), after reviewing various methods for grading evidence, concluded that no existing grading system adequately captured the nuances in interpreting clinical studies and that the use of the available grading systems for rating evidence could increase the risk of recommending


interventions with weak efficacy, even though they may have higher levels of evidence (LoE) (Bandelow et al. 2008). In order to avoid such shortcomings, the WFSBP decided to develop an optimised grading system of specific Levels of Evidence (LoE) for use in development of the WFSBP and similar guidelines. This new system integrated suggestions from other guidelines and used definitions that could be more optimally adapted to the situation of evidential data in psychiatry, in order to provide best transparency for the users of the WFSBP and similar guidelines (Bandelow et al. 2008). In contrast to many other grading systems, the WFSBP gives preference to a thorough analysis of individual studies than to meta-analyses.

Treatment guidelines are of the utmost importance to guide clinicians in their treatment decisions. Because, with an increasing number of possible

CONTACT Alkomiet Hasan  alkomiet.hasan@med.uni-muenchen.de  Department of Psychiatry and Psychotherapy, Klinikum der Universität München, Ludwig-Maximilians Universität München, Nußbaumstraße 7, 80336 München

*Both authors contributed equally to this work.

[†]Michael Bauer (Germany), Michael Berk (Australia), Marijana Bras (Croatia), Philippe Courtet (France), Marijana Bras (Croatia), Bruno Dubois (France), Chin B Eap (Switzerland) Wolfgang Gaebel (Germany), Angelos Halaris (USA), Siegfried Kasper (Austria), Nick Kates (Canada), Walter Kaye (USA), Sidney Kennedy (Canada), Henry R. Kranzler (USA), Rupert Lanzenberger (Austria), Jeffrey Lieberman (USA), Joel Paris (Canada), Georgios Petrides (USA), Dan Rujescu (Germany), Thomas Schlöpfer (Germany), Andrea Schmitt (Germany), Leo Sher (USA), Constantin Soldatos (Greece), Nikos Stefanis (Greece), Florence Thibaut (France), Tanelki Tolga (USA), Janet Treasure (UK), Josef Zohar (Israel).

 Supplemental data for this article can be accessed [here](#).

treatment options, there is an increasing need to define recommendations for endpoints that have not been extensively studied and an increasing discussion about the risk-benefit ratio of available treatments in psychiatry. In that sense, guidelines not only summarise and condense the current medical evidence but also weigh risk-benefit ratios in order to arrive at clinical recommendations. Moreover, guideline developers are aware that performing a clinical study is a difficult work. The goal of grading evidence is not to criticise clinical researchers for the weaknesses of their studies but to help to provide recommendations of the best possible treatment modalities for each patient.

The main objective of this publication is to propose a new evidence and recommendation grading system for use in development of future WFSBP and like treatment guidelines. This paper reviews sources of data typically used in guideline development, potential biases involved in using such data for grading of evidence and will provide guidance on how to use such information in arriving at final grading of evidence and treatment recommendations.

2. Sources of information and dimensions of quality (risk of bias assessment)

Guideline recommendations can be based on (1) systematic search and reviews of original treatment trials (2), meta-analyses, or (3) a synthesis of previously published guidelines. Low quality guidelines could consist of an unstructured conglomerate of a selective choice of open and controlled original studies, regardless of their quality, previously published systematic reviews and/or meta-analyses, previous guidelines and expert opinions which are not based on empirical studies ('eminence-based medicine') (Isaacs and Fitzgerald 1999). In contrast, a high quality guideline ('evidence-based medicine') would be based on an up-to-date and systematic search of available randomised controlled trials (RCT) and meta-analyses, including a quality control of these publications, and in certain scenarios the self-conduct of meta-analyses. Expert opinions would only be used in treatment decisions where no controlled data are available (Guyatt, Oxman, Kunz, et al. 2008; Guyatt, Oxman, Vist, et al. 2008; AWMF 2012; Andrews et al. 2013; Berkman et al. 2015; AWMF 2016, 2017). The dimensions discussed in the next section have been extracted from various sources (Atkins et al. 2004; Bandelow et al. 2008; Guyatt, Oxman, Vist, et al. 2008; AWMF 2012; Andrews et al. 2013; SIGN 2014; Berkman et al. 2015; AWMF

2016, 2017; GRADE 2017) and were set into perspective regarding selected aspects that are important for the field of psychiatry and psychotherapy.

Both clinical trials and meta-analyses have their pros and cons when used as sources to develop guidelines (Huf et al. 2011; AWMF 2012; Greco et al. 2013; da Costa and Juni 2014; AWMF 2017). In the past, the WFSBP grading categories emphasised the importance of well-conducted RCTs for the highest LoE and deemphasized the role of meta-analyses in this process due to a number of methodological reasons (Bandelow et al. 2008). The following paragraphs describe potential and frequent risks-of-bias in clinical trials and meta-analyses. While developing the recommendation grades, the following items should be used for the risk-of-bias assessment (extended according to (AWMF 2012; SIGN 2013, 2014, 2015; AWMF 2016, 2017; NICE TNifHaCE 2017).

Randomised controlled trials

The quality of studies can be checked by different tools. The SIGN statement (SIGN 2014, 2015) is one of the most frequently applied checklists. In general, the following quality aspects have to be considered.

Validity

Methodological publications in this context define two different dimensions of validity of clinical trials, *internal* and *external validity*. Internal validity is defined as the 'extent to which systematic error (bias) is minimized in clinical trials'. Dimensions of internal validity include selection bias, performance bias, detection bias and attrition bias (Juni et al. 2001). External validity is defined as the 'extent to which results of trials provide a correct basis for generalisation to other circumstances' (Juni et al. 2001). Dimensions of external validity include the choice of patients (e.g., stratified by age, sex, severity of disease, comorbidities), the treatment regimen (e.g., stratified by dosage, timing and route of administration, type of treatment with certain treatment classes or concomitant treatments, the setting (primary to tertiary care, outpatient or inpatient treatment, specialisation of care provider), variance in randomisation and the modalities of outcomes (type or definition of outcomes and duration of follow-up) (Juni et al. 2001). External validity can be challenged when the sample is restricted to certain subgroups, e.g., when only patients of a certain age, only male patients, or only patients without comorbidity are included in a trial. The results may not be

generalisable to all patients with the disorder. Other potential methodological limitations of clinical trials that challenge internal and external validity include the study quality (e.g., sample sizes, inadequate power calculations), lack of pre-defined a-priori outcome measures and ‘p-hacking’ (performing multiple statistical tests on the data and only reporting the ones that show significant results) or the quality of reporting. External validity is dependent on the internal validity, because if the internal validity of a given trial is poor, the question of its external validity becomes meaningless (Juni et al. 2001; AWMF 2012, 2016; Bruns and Ioannidis 2016).

Control group

In RCTs, different kinds of control groups can be included. Medication is typically compared to placebo or to an established reference drug. Importantly, it is an area of controversy whether the placebo response has increased over the years (Agid et al. 2013; Bandelow et al. 2015; Khan et al. 2017; Furukawa et al. 2018; Leucht et al. 2018). This phenomenon of an increase in placebo response could be explained to a certain extent by study designs and patient-related features like short trial duration, ‘baseline rating inflation’, heightened expectations of clinicians and patients, the improvement in mental healthcare, flexible-dose treatment designs or the increase in the number of study sites (Agid et al. 2013; Bandelow et al. 2015; Khan et al. 2017; Furukawa et al. 2018; Leucht et al. 2018). This may result in an underestimation of the true drug effect. When comparing to a reference drug, larger sample sizes are needed to demonstrate that the new drug is statistically not inferior to an established drug. In psychotherapy research, waiting lists have been extensively used in the past as a control group. It has been argued that waiting list groups are not an adequate control because such designs cannot exclude the possibility of interpersonal contact as a major contributor to the efficacy when psychotherapeutic intervention is superior. Some authors have argued that waiting lists are unethical, have a high risk for an expectation bias and should be considered to be nocebos rather than placebos, e.g., for anxiety disorder research (Bandelow et al. 2015; Patterson et al. 2016).

Other control group designs include psychological placebos and ‘treatment as usual’ (TAU). However, the definition of TAU may be arbitrary, if we consider that most treatments in clinical practice adhere to some extent to guideline recommendations. The use of active psychological placebos, e.g., duration- and

intensity-matched non-specific treatment such as befriending, attention control, relaxation intervention, supportive counselling (Bendall et al. 2006; Patterson et al. 2016) are probably the only adequate control groups for psychotherapy research, and such approaches are increasingly used. However, manuals for standardised psychological placebo controls should be developed and published to make studies comparable (e.g., Bendall et al. 2003).

Uncontrolled studies

Some authors argue that RCTs do not reflect clinical reality (‘real world’) because patients are highly selected. As an alternative, these authors often suggest the use of naturalistic studies. Often, the same authors advocate the use of treatments that have not been shown to be effective in RCTs. However, naturalistic or open studies are subject to a number of confounding factors, including selection bias, expectancy, allegiance and placebo effects, spontaneous remission, tendency of regression to the mean, lack of control for concomitant treatments and lack of intent-to-treat analysis. Therefore, their scientific value is limited, and therefore they are graded at a much lower LoE.

Uncontrolled studies can only be considered ethical when they are conducted as small scale feasibility or pilot trials of new treatments, in particular for rare disorders for which it is not possible to recruit a sample size large enough for a controlled trial. Regression to the mean and placebo effects mean that many of these risk producing spurious false-positive findings. However, uncontrolled studies, case series and case reports may be helpful for providing guidance on management of treatment refractory patients, e.g., the use of electroconvulsive therapy (ECT) for treatment of psychosis in clozapine-resistant schizophrenia patients or other specific issues such as treatment of specific personality disorders or paraphilias. Moreover, e.g., in bipolar disorder, guideline developers are faced with the problem that relapse prevention trials that show that a certain treatment can prevent depressed and manic episodes over many years are scarce. In these cases, observational or register studies can be used to suggest candidates for future research and treatment recommendations.

Randomisation

It is very rare to find controlled studies that do not use randomisation. However, in some studies, the lack of randomisation is an important source of bias. A typical example of selection bias would be when a pharmacological treatment is compared to a

pharmacological treatment plus psychoeducation, and the patients are not allocated randomly, where an investigator might assign the more severe patients to the combination treatment and the less severe to the drugs only treatment – with the possible result that no differences are found between the two treatment strategies. Beyond the rare issue of no randomisation, failure to report the randomisation procedure is a common methodical limitation, which has to be considered as potential bias.

Blinding

For randomised controlled drug trials, double-blinding is routine, but blinding can also be a potential source of bias. For example, if the studied drug in a blinded trial results in a certain side effect, investigators are at risk to be unblinded (e.g., when the active drug is unveiled by a characteristic side effect). In general, well-conducted blinded clinical trials should undertake every possible step to reduce the risk of accidental unblinding. For other treatments, like ECT or psychological therapies, single blinding can only be achieved by using a ‘blind’ rater, who is blind to the patient’s allocation when assessing the endpoints. However, due to organisational problems, this kind of blinding is at higher risk for unblinding, compared to placebo-controlled trials – unavoidable in psychosocial and lifestyle trials. More advanced rater-blinding techniques may include central raters (e.g., rating via video conference, travelling raters) rather than local raters and the implementation of a dual control method with two independent raters. Moreover, it is recommended that the study statistician is blind to treatment allocation until analysis is complete (‘triple blinding’).

Same conditions for the active and the control group

In most pharmacotherapy trials (especially in registration trials), all conditions should be the same in both the active and the control group, with the only exception of the contents of the study pill. However, in other treatment trials, the surrounding conditions may differ. For example, in psychotherapy and lifestyle modification trials, a comparison can be biased when the length of the treatment group is longer than the control group, e.g., the wait list. Control conditions matched for duration and intensity are recommended. Also, in many psychotherapy trials, the additional naturalistic use of psychopharmacological agents is allowed. The same limitations have to be considered in ECT trials or trials using modern neurostimulation

techniques, like repetitive transcranial magnetic stimulation.

Sample sizes

The sample size has to be adequate, and an a priori sample size calculation for a given endpoint should be provided. Sample sizes should be neither too large nor too small. For example, when a drug is compared to placebo using an extremely large sample, a statistical difference may be found even when the magnitude of difference is very small and not clinically meaningful for the patient. Moreover, very large trials risk recruiting problems, increasing the likelihood that patients with the wrong diagnoses or patients with minimal symptomatic burden are included. Comorbidity can cause similar problems – just because a patient has, say, bipolar disorder, it does not mean that this is the cause of the current symptoms. Substance abuse, for example, may be the dominant operative problem. This can lead to large heterogeneity in a trial that may bias the true difference between the study arms. However, the risk for type II error that is dependent on the sample sizes is a much more common problem in clinical research. For example, when two treatments are compared in a small study without sufficient statistical power, it is often difficult to demonstrate a statistically significant difference, even if a meaningful difference existed. In such cases, it is often erroneously concluded that the new experimental treatment is not better than placebo or, in a comparator trial, is as effective as the established treatment which in fact may not be the case.

Intent-to treat analysis

During standard RCTs in psychiatry, it is estimated that at least one-third or even more of the patients are at risk for drop-out during the study. If only those patients who remained in the study until the endpoint are included in the analysis, the efficacy of the intervention may be overestimated, when patients who dropped out, e.g., due to side effects or to limited efficacy, are not counted. Moreover, one could assume that those patients who remain in a trial represent a positively selected population that is not representative of clinical practice. To control for attrition bias, an intention-to-treat analysis, which evaluates all patients and not only the ones who completed the study (per protocol sample) should be performed (AWMF 2016).

Endpoints

Stakeholders often demand that a new treatment should not only demonstrate symptom reduction and tolerability but also and improvement in subjective wellbeing, quality of life or cost-effectiveness. All medical fields are faced with this need to develop recommendations for such *soft endpoints*, but nearly all randomised controlled trials so far have been conducted with *hard endpoints* like symptomatic improvement, remission, hospitalisation or study discontinuation. Improvement in patients suffering from psychiatric disorders is mostly measured by using rating scales. These can be broadly divided into:

- Symptom-specific scales, e.g., the Positive and Negative Syndrome Scale (PANSS), Young Mania Rating Scale (YMRS) or Hamilton Depression Scale (HAMD)
- Clinical global impression scales, e.g., the Clinical Global Impression (CGI) and Patient Global Impression of improvement (PGI)
- Quality of life scales, e.g., World Health Organisation Quality of Life (WHOQOL) scale, Quality of Life Enjoyment and Satisfaction Questionnaire (Q-Les-Q)

These scales are either used to define a continuous outcome (e.g., decrease in PANSS total) or a dichotomous outcome (e.g., remission according to the Remission in Schizophrenia Working Group criteria (Andreasen et al. 2005) based on the PANSS). Other frequently used outcomes in clinical trials in psychiatry include all-cause discontinuation, which has been used, for example, in the large schizophrenia effectiveness trials (Lieberman et al. 2005; Kahn et al. 2008), or e.g.,

- Numbers of rehospitalisation, length of hospitalisation, study discontinuation, treatment observance, etc.
- Numbers of suicide attempts or suicides
- Cognitive performance
- Personality traits such as impulsivity, etc.
- Metabolic parameters, such as BMI, fasting glucose, or number of cigarettes

As outlined above and as evident from everyday clinical work, it is often demanded that a treatment not only improves symptoms, but also has effects on other outcome dimensions, such as quality of life or health economics (Vos et al. 2005). As an example from the field of cancer research, some treatments for

cancer could significantly prolong life, but are associated with a number of intolerable side effects – such a drug would score high on a symptomatic improvement scale, but low on a quality of life scale, and may not be recommended for use.

However, in a standard RCT, the sample size calculation is often based only on symptomatic improvement. The required simple size of a study depends on the magnitude of the expected effect size difference between the active drug and placebo or between two active drugs. While it is often possible to demonstrate a difference relative to placebo on a symptom specific scale, this is more difficult when it comes to quality of life, where the expected magnitude of effect size is generally lower, and benefits are slower to accrue. In simple words, when a patient suffering from schizophrenia has received treatment for 8 weeks in a RCT, he may experience relief of his paranoid symptoms, but he will very likely not have a new job, a new girlfriend and a new apartment. Thus, he will show improvement on a symptomatic scale, but not on a quality of life scale. Therefore, in most trials showing the efficacy of a certain drug, a difference versus placebo is only found on a symptomatic scale but not on the quality of life scale.

To show the effectiveness of some treatment on quality of life, typically, much larger patient numbers and long follow-up periods are needed. Such studies would be very costly and almost impossible to conduct because it would be hard to recruit enough patients for such a study. This has to be considered when recommendations are developed for soft clinical endpoints. In general, since many endpoints in psychiatry (e.g., improvement in HAMD, MADRS or PANSS) are at higher risk of bias compared to harder endpoints in other areas of medicine (e.g., cardiac events or fasting glucose) because of inter-rater reliability issues. Other endpoints like ‘all cause discontinuation’ may be subject to lower risk of bias and some argue that they may represent the effectiveness in real world clinical practice. Secondary endpoints are always subjected to risk of bias as they are not corrected for multiplicity.

Statistics

As indicated above, there are several different ways to evaluate improvement, including

- Differences in the mean scores of rating scale (e.g., HAMD or PANSS)

- Response (e.g., as defined by as 50% reduction on the HAMD or 25% reduction on the PANSS) (Moller 2008; Leucht 2014)
- Remission (e.g., as defined by as score of seven or less on the HAMD, or as defined by the RSWG criteria using PANSS) (Andreasen et al. 2005; Moller 2008; Leucht 2014)
- Number needed to treat (NNT),
- Survival curves
- Logistic regression analysis
- Other

For patients and clinicians, efficacy expressed as 'response' is easier to understand than a difference in mean scores (e.g., a patient would rather like to hear: 'this drug helped 85% of the patients' than 'this drug showed a significant difference to placebo in end-points scores of 4.3 plus minus 8.6 standard deviation on a rating scale'). For a statistician, both approaches have pros and cons. Mean score differences are usually considered as having a higher statistical power as they are based on continuous measures rather than on more-or-less arbitrary categorical definitions. Data in psychiatry are usually on an *ordinal* scale level (e.g., 3 = severe, 2 = moderate, and 1 = mild). When using dichotomised measures for treatment success (e.g., response, remission or NNT), the scale level goes down to *nominal* level, with the consequence that only tests with lower statistical power can be used.

Sponsor and allegiance effects

Allegiance effects may influence study results. In pharmacological research, study sponsoring by the manufacturer of a certain drug is the most prominent example for allegiance effects. For example, this has been, e.g., shown for antipsychotic trials (e.g., Heres et al. 2006) and for epidemiology trials in neurodegenerative disorders, where industry affiliation has been showed to bias findings (e.g., Cataldo et al. 2010). Especially for new and costly drugs, an influence of the sponsor or manufacturer can be expected. Allegiance effects also have to be considered in psychotherapy research, e.g., when the investigators are advocates of certain psychotherapeutic procedures, or when the investigator is aligned or partnered with the developer of the psychotherapeutic method or when authors have written treatment manuals (Dragioti et al. 2015).

Meta-analyses. Meta-analyses carry certain risks of bias with respect to the selection and quality of included trials, the strategy of data extraction, the

evaluation of the bias within the source data and the applied statistical models or all dimensions of external validity (Huf et al. 2011; da Costa and Juni 2014; AWMF 2017). All these aspects need to be considered when evidence-based recommendations are derived from meta-analyses during the development of treatment guidelines.

Meta-analyses have specific advantages (Fagard et al. 1996; Walker et al. 2008; Stone and Rosopa 2017):

- The results of studies using different scales can be easily compared
- If conflicting results exist with one treatment, i.e., some studies show superiority to placebo and others not, meta-analyses can provide an easy-to-interpret synthesis of all studies by weighing the different studies by sample size and heterogeneity
- When a number of underpowered studies exist, statistical power can be increased by combining these studies in order to receive more reliable results
- Studies that are outliers can be identified and be managed in the analysis
- Moderators of treatment effects can be analysed
- The relative efficacy of various treatment options can be compared in network meta-analyses or pre-post effect size meta-analyses
- Underreporting of small negative studies may be detected by various methods (e.g., funnel plots)

For example, in a comprehensive meta-analysis of all available studies for anxiety disorders, it was shown that large effect size differences exist between various drugs and that medication was more effective than psychotherapy, a result which would not have been found when only looking at the few available head-to-head comparisons (Bandelow et al. 2015).

On the other hand, meta-analyses face specific problems (Sharpe 1997; Stone and Rosopa 2017), which include:

- When there are only small-size RCTs for one treatment, a meta-analysis may over- or underestimate a treatment effect.
- The choice of studies may be subject to arbitrariness in adaption of inclusion and exclusion criteria
- If not controlled via a sensitivity analyses, studies of good and poor quality or studies from heterogeneous samples may be mixed
- When studies are combined that already have sufficient statistical power, the statistical power may be

artificially increased resulting in findings that are a significant but not clinically relevant

- Despite providing statistical approaches (e.g., funnel plot analyses), publication bias can affect the results of meta-analyses as it can affect the development of recommendation based on clinical trials

For example, in one meta-analysis of five studies on lamotrigine for bipolar depression, an overall effect of lamotrigine was found (Geddes et al. 2009). However, four of the five included studies showed no difference to placebo. A systematic review of the five studies would be more likely to come to the conclusion that lamotrigine is not effective in bipolar depression. Thus, the decision how to grade such findings (e.g., by discussing the statistical power of the individual studies versus the pooled results that could result in an overestimation of the effects) must be part of a comprehensive review process as detailed below.

Another important question is whether the pooled drug-class results reported from meta-analyses can be generalised. As an example, one recent meta-analysis indicates the superiority of adding antidepressants or selective serotonin reuptake inhibitors (SSRIs) as a group to an ongoing antipsychotic treatment for depressive or negative symptoms in schizophrenia (Helfer et al. 2016), but can this result be generalised for every SSRI or non-SSRI antidepressants? Were all included trials designed having depressive or negative symptoms as primary outcome parameters? How about the differences between pronounced depressive symptoms and post-psychotic depression? The authors of this well-conducted meta-analysis paid special attention to these issues by providing comprehensive and planned subgroup and sensitivity analyses (Helfer et al. 2016), which reduce the risk of bias, but without this additional information, this meta-analysis would have had a high risk of bias resulting in possibly misleading conclusions. In this context, meta-analysis was used for group of drugs rather than for a single drug, e.g., showing the efficacy of SSRIs in depression. However, it is still difficult to generalise the findings with four SSRIs to a fifth SSRI or a mechanistically different antidepressant. At the same time, one should note that relevant differences in the efficacy of SSRIs, so also for other drug-classes, are difficult to be detected. In the guideline, if possible it is necessary to define which single drugs are effective and not which group of drugs is effective. Therefore, a potential risk for over-generalization in

terms of drug-classes must be addressed. Results from meta-analytic drug-class analyses should be used for recommendations in cases where a differentiation between certain compounds is not possible with sufficient evidence.

Another potential source of bias related to meta-analysis is that in most trials different rating scales are used. Meta-analysis converts the scores of different scales to one effect size, e.g., Cohen's *d*, such as British pounds sterling and US dollars can be converted to euros, thus making prices more comparable. However, in reality the effect sizes from the same study with the same patients may differ considerably when different scales are used. For example, when in one trial on MDD two rating scales are used, e.g., the Hamilton Depression Rating Scale (HAMD) and the Montgomery-Åsberg Depression Rating Scale (MADRS), both scales should yield exactly the same effect size. However, in clinical trials, two scales can differ widely although they claim to measure the same construct. The reason for such discrepancy could be related to the composition of the different items, e.g., three items for sleep in one scale (HAMD) and one item for sleep in the other scale (MADRS). Another source of bias in this context is the use of different version of a scale (e.g., HAMD-17, HAMD-21 or HAMD-24) for the same clinical research question.

One should note that the number of published meta-analyses show a virtually exponential increase in the last years, whereas RCTs are not published at this rapidity (da Costa and Juni 2014). This development highlights the need to take into account all available sources of evidence to reduce the time-lag between publication of clinical trials and the development of clinical treatment recommendations. In summary, meta-analyses need, as individual trials, to undergo a strict quality control if the results are used for guideline recommendations. In reality, two meta-analyses exploring the same data, e.g., antidepressants for bipolar disorders, can provide contrasting recommendations, based on selection criteria and date of conduct (Sidor and Macqueen 2011; McGirr et al. 2016). As is mandatory for clinical trials, the protocols of meta-analysis (including the planned meta-analytic models) should be published (e.g., at PROSPERO) prior to the systematic literature search to reduce the risk of selection bias. Several tools (e.g., PRISMA (Moher et al. 2009), AMSTAR (Shea et al. 2009, 2017) or ROBIS (Whiting et al. 2016)) are available to evaluate the quality and the risks of bias in meta-analyses and systematic reviews.

3. General aspects of grading evidence

As can be gathered from the above review, the risk-of-bias assessment is the most important step in evaluating the sources of evidence used for a guideline, irrespective of whether clinical trials or meta-analysis or systematic reviews are used. From the WFSBP perspective, there should be a step-by-step approach when grading evidence:

1. Prioritise and evaluate (risk-of-bias assessment) single RCTs: when sufficient RCTs exist for a certain treatment and these are of high quality and do not contradict each other, this approach is preferred.
2. Evaluate meta-analyses (risk-of-bias assessment): when there are at least three RCTs for one treatment and these are inconsistent – meaning that some studies show a difference to placebo and others do not – meta-analyses of high quality should be used.
3. Evaluate systematic reviews without meta-analysis (risk-of-bias assessment). This source of evidence should only be used if no recommendations can be generated from (1) and (2).
4. From our viewpoint we do not recommend to base the evidence grading on non-systematic reviews.

Other grading systems

We have qualitatively summarised different available and frequently used grading systems (see [Supplementary Data](#)). As outlined above, many grading systems are available, but most of the currently used systems are related to the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system (Andrews et al. 2013; GRADE 2017). GRADE has been defined as a systematic and explicit approach to making judgements about quality of evidence and strength of recommendations (Atkins et al. 2004; Guyatt, Oxman, Kunz, et al. 2008; Guyatt, Oxman, Vist, et al. 2008) and the main aspects of GRADE correspond to the methodology of original research, to the consistency across studies, to the generalizability of result and to the efficacy of a given treatment (GRADE 2017). These aspects are the core of many available grading systems and of those used in the past. [Supplementary Table S1](#) provides an overview of different available grading systems and the details are displayed in supplementary Tables S2 to S8. Certain elements from these grading systems have been adapted and modified for the new WFSBP grading system taking into account the available WFSBP

system (Bandelow et al. 2008). Every grading system has its advantages and disadvantages, and one could assume that no grading system is perfect. From the perspective of the WFSBP guideline development, the following issues may limit the applicability of a given grading system:

- The strict prioritisation of meta-analyses compared to reviewing individual clinical trials, without an evaluation whether both sources of evidence have contradicting results
- The lack of guidance regarding how to deal with conflicting results, especially when uncontrolled studies are used for developing treatment recommendations
- The lack of differentiation regarding ‘absence of evidence of an effect’ (lack of efficacy) and ‘evidence of absence of an effect’ (lack of trial data)
- The difficulty in differentiating a true negative trial from a failed trial (usually for methodological reasons)
- The lack of possibility to give negative recommendations (e.g., when RCTs have shown evidence of non-efficacy or in cases in which treatments should not be used because of risk of severe adverse effects)

Specific differences between the systems are the number of LoE steps and how the LoE is translated to recommendation grades. Moreover, the systems differ in the fact whether RCTs or systematic reviews/meta-analyses are needed to reach the highest LoE. In the various grading systems, the number of LoE was found to be between 3 and 10 and the number of GoR was between 3 and 5.

While more levels might provide more precision, they might add more complexity and might be less useful from clinical practice perspective, thus defeating the purpose of clinical practice guidelines. The WFSBP believes that an estimated average of 3+1 LoE (A, B, C and D (no evidence)) and 3+1 recommendation grades (1, 2, 3 and no recommendation possible grade) might achieve that balance as such has the potential to adequately capture the main information from all systems and provide recommendations that might be easier for adaptation in clinical practice (see [Tables 1](#) and [2](#)).

4. Proposed new WFSBP grading system

The proposed new WFSBP grading system is displayed in [Tables 1](#) and [2](#) and is planned for all future WFSBP

Table 1. New WFSBP grading system (levels of evidence).

Levels of evidence (LoE)					
Evidence that the intervention is effective	Grade	Explanation	Evidence that the intervention is NOT effective	Grade	Explanation
<i>Strong</i>	A	At least two independent RCTs with a low risk of bias show efficacy (superiority to placebo, or in the case of psychotherapy studies, superiority to a 'active psychological placebo' in a study with adequate blinding), OR superiority to/equivalent efficacy compared with an established comparator treatment in a three-arm study with placebo control or in a well-powered non-inferiority trial (only applicable if such a standard treatment exists) with a low risk of bias, AND No negative RCTs with a low risk of bias exist. If there are contradicting results from RCTs, the majority of RCTs AND/OR a meta-analysis with low risk of bias shows efficacy. If there are more than one 'A' treatment options, the decision should be based on head-to-head comparisons or meta-analyses showing superiority of one of the treatments	<i>Strong</i>	-A	At least two independent, adequately powered RCTs with low risk of bias as detailed on the left show NO efficacy AND no positive RCTs with a low risk of bias exist If there are contradicting results from RCTs, however, the majority of RCTs AND/OR a meta-analysis with very low risk of bias shows NO efficacy
<i>Limited</i>	B	One RCT with a moderate risk of bias showing superiority to placebo (or in the case of psychotherapy studies, superiority to a 'active psychological placebo') OR A randomised controlled comparison with a standard treatment without placebo control with a sample size sufficient for a non-inferiority trial with a moderate risk of bias, AND No negative studies exist OR Meta-analyses with a moderate risk of bias that show efficacy	<i>Limited</i>	-B	One RCT with a moderate risk of bias showing NO superiority to placebo (or in the case of psychotherapy studies, NO superiority to a 'active psychological placebo') OR LESS efficacy than a standard treatment in an RCT without placebo control with a moderate risk of bias AND No positive studies exist OR Meta-analyses with a moderate risk of bias that show NO efficacy
<i>Low</i>	C1	One or more prospective open studies (with a minimum of 10 evaluable patients per group) using a control group, but no randomisation, or using no control group, show efficacy. OR One or more well-conducted case control or cohort studies (with a minimum of 10 evaluable patients) with a moderate probability that the relationship is causal show efficacy OR RCTs AND/OR meta-analyses with a high risk of bias show efficacy	<i>Low</i>	-C1	One or more prospective open studies (with a minimum of 10 evaluable patients) using a control group, but no randomisation, or using no control group, show NO efficacy. OR One or more well-conducted case control or cohort studies (with a minimum of 10 evaluable patients) with a moderate probability that the relationship is causal shows NO efficacy OR RCTs AND/OR meta-Analyses with a high risk of bias that show NO efficacy
	C2	Non-analytic studies, e.g., case reports or case series with less than 10 evaluable patients show efficacy in the majority of cases		-C2	Non-analytic studies, e.g., case reports or case series with less than 10 evaluable patients show NO efficacy in the majority of cases
	C3	Expert opinions not based on any published data reporting efficacy		-C3	Expert opinions not based on any published data reporting NO efficacy
<i>No Evidence</i>	D	No sufficient evidence to advise for or against the use of the intervention			

and related guideline revisions. The following paragraphs describe the new system and how it can be used to develop WFSBP guidelines.

The revised WFSBP grading system incorporates the three following principles:

1. The systems accepts clinical trials, meta-analyses as well as cohort studies from national or international registers for grading. However, clinical

trials are still prioritised as per the last version of this grading system (Bandelow et al. 2008). The results of clinical trials can be corroborated by a meta-analysis (this would enable calculation of the effect size) as described in Table 1.

2. The number of LoE and GoR should be limited to a 3 + 1 system to allow a pragmatic application in clinical practice

Table 2. New WFSBP grading system (Grades of recommendation).

Grades of recommendation (GoR)

GoR based on a synthesis of:

- I. Level of evidence (see Table 1)
- II. Acceptability (criteria for grading adapted and modified from (AWMF 2012; GRADE 2017)), rated 'strong', 'limited' and 'weak'
 - Risk–benefit ratio (e.g., adverse effects, interactions)
 - Cost–benefit ratio
 - Applicability in the target population
 - Ethical and legal aspects
 - Preferences of service users
 - Practicability

The algorithm of evaluating the acceptability to develop GoR from LoE is detailed below. For negative recommendations, LoE can be directly translated into GoR as also detailed below.

Recommendation for using the intervention	Grade		Recommendation AGAINST using the intervention	Grade	
<i>Strong</i>	1	'A' LoE and GOOD acceptability	<i>Strong</i>	–1	Strong negative evidence (LoE -A)
<i>Limited</i>	2	'A' LoE and MODERATE acceptability	<i>Limited</i>	–2	Limited negative evidence (LoE -B)
<i>Weak</i>	3	OR 'B' LoE and GOOD acceptability 'A' LoE and POOR acceptability OR 'B' LoE and MODERATE/POOR acceptability OR 'C' LoE and GOOD/MODERATE/POOR acceptability	<i>Weak</i>	–3	Weak negative evidence (LoE -C)
<i>No recommendation possible</i>	4		Insufficient evidence (LoE D) to give recommendations		

3. A separation of LoE and GoR is needed to allow to define first, second, third, etc., lines of treatment based on the quality of the source data, risk-benefit evaluation and other criteria for grading recommendations as detailed below.

Following these specifications, we aimed at defining a system on the basis of the available grading systems with adaptations wherever necessary. In this context, the Association of the Scientific Medical Societies in Germany (AWMF; <http://www.awmf.org/leitlinien/awmf-regelwerk/II-entwicklung.html>.) developed a GRADE-based (GRADE 2017) procedure for a structured consensus process and defined how LoE can be translated to GoR with clinical relevant gradings (AWMF 2012). We used and modified these available criteria for grading (AWMF 2012), to develop a grading of acceptability (see Table 2).

The WFSBP Guidelines henceforth should follow the following :

1. **WFSBP Guidelines:** all future WFSBP guidelines should follow the new WFSBP grading System.
2. **WFSBP guideline committee:** the chair of each WFSBP Task Force or the first author of the respective guideline identify a lead and the core group for leading the development of a guideline

for a particular disorder in consultation with the members of that Task Force. The core group should include members who have substantial clinical experience with the disorder and/or have published relevant peer-reviewed papers on this disorder. All members have to declare financial and non-financial conflicts of interest, otherwise a contribution as an author or as a task force member on the final publication is not possible. When a voting is held on a certain treatment, members who have a conflict of interest with regard to this treatment should be excluded from the referendum. Further, all members of the Task Force need to be provided with the opportunity to contribute and the authorship is based on the contribution and participation.

3. **Limit the number of LoE and GoR to 3 + 1 for each.** The new system has now three LoE and GoR plus one LoE for 'no evidence' and one grade for 'no recommendation possible'. This new system adapts suggestions from various available grading systems and the previous WFSBP grading system (Bandelow et al. 2008).
4. **Clarify recommendation grades for negative evidence.** In cases where the majority of RCTs show non-superiority to placebo (or in the case of psychotherapy studies, non-superiority to a

'psychological placebo') or inferiority to comparator treatment and an available meta-analysis is also negative (or no meta-analysis is available), negative evidence can be defined as detailed in [Table 1](#). If negative evidence from RCTs contradicts a positive meta-analysis of the same methodological quality or vice versa, the recommendations should be extrapolated by downgrading the recommendation based on the criteria detailed in [Table 2](#). If the methodological qualities differ, the source with the higher level of quality should be used for developing the treatment recommendation.

5. **Keep a recommendation grade for the lack of evidence.** In cases where no evidence is available to balance risks versus benefits, the recommendation grade 4 (insufficient evidence, see [Table 1](#)) should be used (please see Supplementary tables).
6. **Develop a system that allows to up- or downgrade recommendations** based on the criteria for grading recommendations detailed below as suggested by AWMF (AWMF 2012) and GRADE (Berkman et al. 2015; GRADE 2017). While LoE has to be developed using the procedures displayed in [Table 1](#) as a result of a strict risk of bias assessment, GoR have to be developed or extrapolated as detailed below and in [Table 2](#). Usually high LoE will result in high recommendation grades, etc., but WFSBP guideline developers can downgrade a high LoE to a low GoR taking into account the criteria for grading and the new acceptability ranks (see [Table 2](#)). From a theoretical view, the same process can be used to upgrade a lower LoE to a higher GoR, but this direction (so-called extrapolated evidence) should only be used in rare cases only and must be clearly explained. For WFSBP guidelines it is not recommended to upgrade GoR with extrapolated evidence, but we wanted to include this possibility for recommendations that we cannot foresee at the moment. The decision should be made after an objective and balanced discussion of the available sources of evidence.

Our proposed and adapted multi-step system based on the available grading systems (AHCPR 1992; Guyatt, Oxman, Vist, et al. 2008; CEBM 2009; NHMRC 2009; Owens et al. 2009; AWMF 2012; USPSTF 2012; Andrews et al. 2013; SIGN 2013; GRADE 2017) to generate LoE that allow both randomised controlled trials

and, under specific conditions, meta-analyses (see [Table 1](#)), to be the basis for the highest evidence and recommendations levels. We extended an available (AWMF 2012) key set of features to be used for this adaptation process between LoE and GoR.

The features (criteria for grading recommendations) should be discussed when developing recommendation grades are (see [Table 2](#)):

Evaluating risks of bias

- Quality of clinical trials/meta-analyses/other source results, precision of effect estimates
- Clinical relevance of (primary outcomes), effect sizes
- Statistical heterogeneity and stratification analyses (e.g., centre effects, effects of a potential bias-inducing subgroup such as age, gender or ethnic groups)
- Other

Acceptability

- Risk–benefit ratio (e.g., adverse effects, interactions)
- Cost–benefit ratio
- Applicability in the target population
- Ethical and legal aspects
- Preferences of service users
- Practicability
- Other

In summary, the process of developing recommendation grades from the available evidence (adapted and extended according to Atkins et al. 2004; Guyatt, Oxman, Vist, et al. 2008; AWMF 2012; Andrews et al. 2013; AWMF 2016, 2017; GRADE 2017; NICE TNiFHaCE 2017) should include the following steps:

1. Define the clinical question that will be answered with a recommendation grade and aim to define an evidence grade based on original clinical trials (see [Table 1](#))
2. Systematically search the literature following, e.g., the PRISMA (Moher et al. 2009) recommendations and/or adapt high-grade guidelines after methodological evaluation and/or use recent meta-analyses/systematic reviews after methodological evaluation (e.g., AMSTAR, ROBIS) (Shea et al. 2009; Whiting et al. 2016; Shea et al. 2017) to provide a comprehensive overview of the available evidence. Evaluate the quality of included publications using, e.g., SIGN checklists (SIGN 2014, 2015) or related tools. As an alternative other available

- high-quality guidelines can be used to identify relevant sources of evidence.
3. If RCTs are used for recommendations, evaluate potential risks of bias, e.g., using the GRADE (Guyatt, Oxman, Vist, et al. 2008) or Cochrane tools (Cochrane Training 2017) and evaluate internal and external validity (e.g., by using the SIGN checklists, SIGN 2014, 2015)
 4. Define the LoE.
 5. Translate the LoE to GoR by taking the aforementioned criteria for grading recommendations (Atkins et al. 2004; Guyatt, Oxman, Kunz, et al. 2008; AWMF 2012; Berkman et al. 2015; GRADE 2017) into consideration (see Table 2). Each recommendation should be accompanied by a methodological discussion of how the LoE were translated to recommendation grades.
 6. Phrase the recommendation in that way that an intervention or a diagnostic procedure is 'offered' to the service user. If negative recommendations have to be phrased, follow the same process as defined for positive recommendations.
 7. Use the suggested wording that has been modified according to the NICE Strength recommendation system (Addington et al. 2017; Crockford and Addington 2017; NICE TNIfHaCE 2017) (please see Supplementary tables) ('must', 'should', 'could', 'may') to phrase and emphasise the respective recommendations (see Table 2, see also Supplementary data). In the new WFSBP system 'should' relates to GoR = 1, 'could' relates to GoR = 2 and 'may' to GoR = 3. The use of 'must' is limited to recommendations if there is a legal duty to apply the recommendation (e.g., '*weekly WBC assessments must be performed in the first 18 weeks of clozapine use*') or for very serious recommendations (e.g., for recommendation were patients may die: '*clozapine must not be combined with intravenous benzodiazepines*') as defined by NICE. Thus, the wording 'must' is not related to any GoR.
 8. Build recommendations based on the PICO (patient/population, intervention, comparison, outcome) clinical question scheme (Richardson et al. 1995) taking into account bullet points 6 and 7 defined above (for example: '*For patients with treatment-resistant schizophrenia (P) clozapine (I) should be offered instead of another antipsychotics (C) to reduce positive symptoms (O)*' – Level of evidence: A (strong), Grade of recommendation: 1 (strong)).

9. Document this process and publish a supplementary document online together with the WFSBP guideline.

Situations where the cumulative evidence from meta-analyses is in contrast to level A RCTs or where only negative studies result in a positive effect in a meta-analysis due to the increased pooled statistical power require a comprehensive discussion. In such constellations, an attenuation of the recommendation grade despite the highest LoE can be used to present the conflicting evidence. Uncontrolled studies should not be used to justify treatment recommendations in situation where higher sources of evidence are available. When level A RCTs are used for single recommendations, a comprehensive discussion of the selection bias risks is needed. If meta-analyses with conflicting results compared to such RCTs are available, an adaptation of the recommendation grade and a careful explanation should also be considered. In situations where conflicting data exist (e.g., one positive versus one negative RCT with the same LoE or positive meta-analyses versus many negative RCTs), a strict and comprehensive methodological discussion and a discussion of the criteria of grading recommendations must be used to explain the phrased recommendation. RCTs and other trials with lower LoE (B, C) should only be used for recommendations where no other source of evidence is available.

5. Discussion

So what is the best way to grade evidence – focussing on randomised trials or on meta-analyses? The truth lies somewhere in between, because we need both sources of evidence to develop treatment guidelines in psychiatry that summarise all available evidence that are relevant for patient care in different sectors of the healthcare systems and that are comparable to guidelines in other fields of medicine. Both approaches have common and different risks of bias and the evaluation and consideration of those risks reduces the risk of unjustified treatment recommendations. However, we still believe that evidence derived from well-conducted randomised controlled trials with a low risk of bias remains the gold standard, keeping in mind that many more such trials are urgently needed. However, performing well-designed clinical trials is costly and needs a lot of financial and non-financial resources. Thus, governments, policymakers, funding institutions and healthcare foundations should set-up specific funding programmes that allow to

conduct industry-independent clinical trials. Moreover, the integration of more gender-sensitive perspectives in all aspects of clinical and preclinical research is needed. Importantly, the European Medicines Agency (EMA) aims at developing separate guidelines for woman as a specific population in clinical trials (Thibaut 2017).

We are aware that no grading system, including our proposed new WFSBP grading system, is perfect and that every available system has potential limitations. Moreover, we recognise that developing guidelines with a high methodological level is time consuming and costly, and that in most countries no sufficient funding for this tremendous work is provided. Thus, combining the work (e.g., systematic literature search, evaluation of source of evidence) for national and WFSBP guidelines (e.g., when the authors of a WFSBP guideline are also responsible for a national guideline), developing guidelines beyond national borders, developing 'living guidelines' that are continuously updated or provide 'focused guidelines' that address only specific and critical issues in the treatment of a given disorders are potential solutions for this evident problem. Finally, 'pragmatic guidelines' with less rigorous methodological standards than described here, may be an alternative solution when resources are limited. However, such compromises should never result in low-quality guidelines that have the character of expert-opinion papers or selective reviews.

In summary, the overall aim of this paper was to define one theoretical path to develop guidelines and we recommend to follow these specifications as far as possible when developing new or updating available WFSBP guidelines. The strict methodological evaluation of all available sources of evidence and the critical appraisal of published positive and negative findings as basis for developing treatment recommendations that have reached consensus will allow to draft guidelines that are scientifically valid and clinically relevant. We believe that evidence-based consensus guidelines will improve the quality and acceptance of treatment and will identify areas in which further high-quality research is needed.

Acknowledgements

The draft version of this manuscript was sent to all WFSBP task force leads and to all presidents of the various national societies of biological psychiatry that are members of the WFSBP for internal review; our thanks are addressed to those presidents who have sent us their comments on the

guidelines. MB is supported by a NHMRC Senior Principal Research Fellowship (1059660 and 1156072).

Disclosure statement

In the last three years, Alkomiet Hasan received paid speakership by Lundbeck, Janssen-Cilag and Otsuka. He was member of advisory boards of Lundbeck, Janssen-Cilag, Roche and Otsuka. In the past 3 years, Borwin Bandelow has been on the speakers' board for Hexal, Janssen, Lilly, and Lundbeck and on the advisory board for Mundipharma. Lakshmi Yatham has received research grants from or has been on speaker advisory boards for Allergan, AstraZeneca, Alkermes, Bristol-Myers Squibb, Canadian Institutes of Health Research, Canadian Network for Mood and Anxiety Treatments, Dainippon Sumitomo Inc, Eli Lilly & Co., Forrest, GlaxoSmithKline, Janssen, Lundbeck, Michael Smith Foundation for Health Research, Novartis, Otsuka, Pfizer, Ranbaxy, Servier, Sunovion, the Stanley Foundation, Teva and Valeant Pharmaceuticals. Michael Berk has received Grant/Research Support from the NIH, Cooperative Research Centre, Simons Autism Foundation, Cancer Council of Victoria, Stanley Medical Research Foundation, MBF, NHMRC, Beyond Blue, Rotary Health, Meat and Livestock Board, AstraZeneca, Woolworths, Avant and the Harry Windsor Foundation, book royalties from Oxford University Press, Cambridge University Press, Springer Nature and Allen and Unwin, has been a speaker for AstraZeneca, Lundbeck, Merck and Servier and served as a consultant to Allergan, AstraZeneca, Bioadvantex, Bionomics, Collaborative Medicinal Development, Grunbionics, Janssen Cilag, LivaNova, Lundbeck, Merck, Mylan, Otsuka and Servier. Peter Falkai has been an honorary speaker for AstraZeneca, Bristol Myers Squibb, Eli Lilly, Essex, GE Healthcare, GlaxoSmithKline, Janssen-Cilag, Lundbeck, Otsuka, Pfizer, Servier and Takeda, and during the past five years, but not presently, has been a member of the advisory boards of Janssen-Cilag, AstraZeneca, Eli Lilly and Lundbeck. Hans-Jürgen Möller received honoraria for lectures or expert meetings by the following pharmaceutical companies: Lundbeck, Schwabe, Otsuka and Servier. Siegfried Kasper received grants/research support, consulting fees and/or honoraria within the last three years from Angelini, AOP Orphan Pharmaceuticals AG, Celegne GmbH, Eli Lilly, Janssen-Cilag Pharma GmbH, KRKA-Pharma, Lundbeck A/S, Mundipharma, Neuraxpharm, Pfizer, Sanofi, Schwabe, Servier, Shire, Sumitomo Dainippon Pharma Co. Ltd. and Takeda.

ORCID

Borwin Bandelow  <http://orcid.org/0000-0003-2511-3768>
Michael Berk  <http://orcid.org/0000-0002-5554-6946>

References

- Addington D, Abidi S, Garcia-Ortega I, Honer WG, Ismail Z. 2017. Canadian guidelines for the assessment and diagnosis of patients with schizophrenia spectrum and other psychotic disorders. *Can J Psychiatry*. 62:594–603.

- Agid O, Siu CO, Potkin SG, Kapur S, Watsky E, Vanderburg D, Zipursky RB, Remington G. 2013. Meta-regression analysis of placebo response in antipsychotic trials, 1970-2010. *AJP*. 170:1335-1344.
- [AHCPR] Agency for Health Care Policy and Research Publications. 1992. Acute pain management: operative or medical procedures and trauma. Rockville (MD).
- Andreasen NC, Carpenter WT, Jr., Kane JM, Lasser RA, Marder SR, Weinberger DR. 2005. Remission in schizophrenia: proposed criteria and rationale for consensus. *AJP*. 162: 441-449.
- Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, Nasser M, Meerpohl J, Post PN, Kunz R, et al. 2013. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol*. 66:719-725.
- Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, et al. 2004. Grading quality of evidence and strength of recommendations. *BMJ*. 328:1490
- [AWMF] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. 2012. AWMF-Regelwerk „Leitlinien“. 1. Auflage. [accessed 2017 December 26]. <http://www.awmf.org/leitlinien/awmf-regelwerk.html> <http://www.awmf.org/leitlinien/awmf-regelwerk/II-entwicklung/awmf-regelwerk-03-leitlinienentwicklung/II-entwicklung-graduierung-der-empfehlungen.html>.
- [AWMF] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. 2016. Bewertung des Biasrisikos (Risiko systematischer Fehler) in klinischen Studien: ein Manual für die Leitlinienerstellung. [accessed 2017 December 26]. Cochrane Germany: <http://www.cochrane.de/de/rob-manual>
- [AWMF] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. 2017. Bewertung von systematischen Übersichtsarbeiten: ein Manual für die Leitlinienerstellung. [accessed 2017 December 26]. Cochrane Germany: <http://www.cochrane.de/de/review-bewertung-manual>
- Bandelow B, Reitt M, Rover C, Michaelis S, Gorlich Y, Wedekind D. 2015. Efficacy of treatments for anxiety disorders: a meta-analysis. *Int Clin Psychopharmacol*. 30: 183-192.
- Bandelow B, Zohar J, Kasper S, Moller HJ. 2008. How to grade categories of evidence. *World J Biol Psychiatry*. 9: 242-247.
- Bendall S, Jackson HJ, Killackey E, Allott K, Johnson T, Harrigan S, Gleeson J, McGorry PD. 2006. The credibility and acceptability of befriending as a control therapy in a randomized controlled trial of cognitive behaviour therapy for acute first episode psychosis. *Behav Cognit Psychother*. 34:277-291.
- Bendall S, Killackey E, Jackson H, Gleeson J. 2003. Befriending manual. Melbourne: ORYGEN Research Centre, University of Melbourne.
- Berkman ND, Lohr KN, Ansari MT, Balk EM, Kane R, McDonagh M, Morton SC, Viswanathan M, Bass EB, Butler M, et al. 2015. Grading the strength of a body of evidence when assessing health care interventions: an EPC update. *J Clin Epidemiol*. 68:1312-1324.
- Bruns SB, Ioannidis JP. 2016. p-Curve and p-Hacking in observational research. *PLoS One*. 11:e0149144.
- Cataldo JK, Prochaska JJ, Glantz SA. 2010. Cigarette smoking is a risk factor for Alzheimer's Disease: an analysis controlling for tobacco industry affiliation. *JAD*. 19:465-480.
- [CEBM] Oxford Centre for Evidence-based Medicine. 2009. [accessed 2017 December 30]. <http://www.cebm.net/blog/2009/06/11/oxford-centre-evidence-based-medicine-level-sevidence-march-2009/>.
- Cochrane Training. 2017. Cochrane Handbook for Systematic Reviews of Interventions. [accessed 2017 December 26].
- Crockford D, Addington D. 2017. Canadian schizophrenia guidelines: schizophrenia and other psychotic disorders with coexisting substance use disorders. *Can J Psychiatry*. 62:624-634.
- da Costa BR, Juni P. 2014. Systematic reviews and meta-analyses of randomized trials: principles and pitfalls. *European Heart J*. 35:3336-3345.
- Dragioti E, Dimoliatis I, Fountoulakis KN, Evangelou E. 2015. A systematic appraisal of allegiance effect in randomized controlled trials of psychotherapy. *Ann Gen Psychiatry*. 14:25.
- Fagard RH, Staessen JA, Thijs L. 1996. Advantages and disadvantages of the meta-analysis approach. *J Hypertension Suppl*. 14:S9-S12, discussion S13.
- Furukawa TA, Cipriani A, Leucht S, Atkinson LZ, Ogawa Y, Takeshima N, Hayasaka Y, Chaimani A, Salanti G. 2018. Is placebo response in antidepressant trials rising or not? A reanalysis of datasets to conclude this long-lasting controversy. *Evid Based Ment Health*. 21:1-3.
- Geddes JR, Calabrese JR, Goodwin GM. 2009. Lamotrigine for treatment of bipolar depression: independent meta-analysis and meta-regression of individual patient data from five randomised trials. *Br J Psychiatry*. 194:4-9.
- GRADE. 2017. Grading of Recommendations Assessment, Development and Evaluation. [accessed 2017 December 30]. <http://gradeworkinggroup.org/>.
- Greco T, Zangrillo A, Biondi-Zoccai G, Landoni G. 2013. Meta-analysis: pitfalls and hints. *Heart Lung Vessel*. 5: 219-225.
- Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schunemann HJ, Group GW. 2008. Going from evidence to recommendations. *BMJ*. 336:1049-1051.
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ, Group GW. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 336: 924-926.
- Helfer B, Samara MT, Huhn M, Klupp E, Leucht C, Zhu Y, Engel RR, Leucht S. 2016. Efficacy and safety of antidepressants added to antipsychotics for schizophrenia: a systematic review and meta-analysis. *AJP*. 173:876-886.
- Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S. 2006. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. *Am J Psychiatry*. 163: 185-194.
- Huf W, Kalcher K, Pail G, Friedrich ME, Filzmoser P, Kasper S. 2011. Meta-analysis: fact or fiction? How to interpret meta-analyses. *World J Biol Psychiatry*. 12:188-200.
- Isaacs D, Fitzgerald D. 1999. Seven alternatives to evidence based medicine. *BMJ*. 319:1618.

- Juni P, Altman DG, Egger M. 2001. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ*. 323:42–46.
- Kahn RS, Fleischhacker WW, Boter H, Davidson M, Vergouwe Y, Keet IP, Gheorghe MD, Rybakowski JK, Galderisi S, Libiger J, et al. 2008. Effectiveness of antipsychotic drugs in first-episode schizophrenia and schizophreniform disorder: an open randomised clinical trial. *Lancet*. 371: 1085–1097.
- Khan A, Fahl Mar K, Faucett J, Khan Schilling S, Brown WA. 2017. Has the rising placebo response impacted antidepressant clinical trial outcome? Data from the US Food and Drug Administration 1987–2013. *World Psychiatry*. 16: 181–192.
- Leucht S. 2014. Measurements of response, remission, and recovery in schizophrenia and examples for their clinical application. *J Clin Psychiatry*. 75: 8–14.
- Leucht S, Chaimani A, Leucht C, Huhn M, Mavridis D, Helfer B, Samara M, Cipriani A, Geddes JR, Salanti G, et al. 2018. 60 years of placebo-controlled antipsychotic drug trials in acute schizophrenia: meta-regression of predictors of placebo response. *Schizophr Res*. 201:315–323.
- Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, Keefe RS, Davis SM, Davis CE, Lebowitz BD, et al. 2005. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N Engl J Med*. 353: 1209–1223.
- McGirr A, Vohringer PA, Ghaemi SN, Lam RW, Yatham LN. 2016. Safety and efficacy of adjunctive second-generation antidepressant therapy with a mood stabiliser or an atypical antipsychotic in acute bipolar depression: a systematic review and meta-analysis of randomised placebo-controlled trials. *Lancet Psychiatry*. 3:1138–1146.
- Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 339:b2535.
- Moller HJ. 2008. Outcomes in major depressive disorder: the evolving concept of remission and its implications for treatment. *World J Biol Psychiatry*. 9:102–114.
- NHMRC. 2009. NHMRC additional levels of evidence and grades for recommendations for developers of guidelines. [accessed 30.12.2017]. https://www.nhmrc.gov.au/_files_nhmrc/file/guidelines/developers/nhmrc_levels_grades_evidence_120423.pdf.
- NICE TNiFHaCE. 2017. Developing NICE guidelines: the manual. [accessed 2017 December 27]. <https://www.nice.org.uk/process/pmg20/chapter/developing-and-wording-recommendations-and-writing-the-guideline>.
- Owens DK, Lohr KN, Atkins D. 2009. Grading the strength of a body of evidence when comparing medical interventions. In: Agency for Healthcare Research and Quality. *Methods Guide for Comparative Effectiveness Reviews*.
- Patterson B, Boyle MH, Kivlenieks M, Van Ameringen M. 2016. The use of waitlists as control conditions in anxiety disorders research. *J Psychiatr Res*. 83:112–120.
- Richardson WS, Wilson MC, Nishikawa J, Hayward RS. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 123:A12–A13.
- Sharpe D. 1997. Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. *Clin Psychol Rev*. 17:881–901.
- Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, Henry DA, Boers M. 2009. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol*. 62: 1013–1020.
- Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, et al. 2017. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 358:j4008.
- Sidor MM, Macqueen GM. 2011. Antidepressants for the acute treatment of bipolar depression: a systematic review and meta-analysis. *J Clin Psychiatry*. 72:156–167.
- [SIGN] Scottish Intercollegiate Guidelines Network. 2013. SIGN 131 • Management of schizophrenia • A national clinical guideline. <https://www.sign.ac.uk/sign-131-management-of-schizophrenia.html>
- [SIGN] Scottish Intercollegiate Guidelines Network. 2014. Critical appraisal notes and checklists. [accessed 27.12.2017]. <http://www.sign.ac.uk/checklists-and-notes.html>.
- [SIGN] Scottish Intercollegiate Guidelines Network. 2015. SIGN 50: a guideline developer's handbook Vol. (SIGN publication no. 50). https://www.sign.ac.uk/assets/sign50_2011.pdf.
- Stone DL, Rosopa PJ. 2017. The advantages and limitations of using meta-analysis in human resource management research. *Human Resource Manage Rev*. 27:1–7.
- Thibaut F. 2017. Gender does matter in clinical research. *Eur Arch Psychiatry Clin Neurosci*. 267:283–284.
- [USPSTF] U.S. Preventive Services Task Force. 2012. U.S. Preventive Services Task Force (USPSTF) grading system. [accessed 2017 December 30]. <https://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions>.
- Vos T, Haby MM, Magnus A, Mihalopoulos C, Andrews G, Carter R. 2005. Assessing cost-effectiveness in mental health: helping policy-makers prioritize and plan health services. *Aust N Z J Psychiatry*. 39:701–712.
- Walker E, Hernandez AV, Kattan MW. 2008. Meta-analysis: its strengths and limitations. *Cleve Clin J Med*. 75:431–439.
- Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R, Group R. 2016. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol*. 69:225–234.