Discussion Forum

Trust, but verify. The errors and misinterpretations in the Cochrane analysis by O. J. Storebo and colleagues on the efficacy and safety of methylphenidate for the treatment of children and adolescents with ADHD

T. Banaschewski¹, M. Gerlach², K. Becker³, M. Holtmann⁴, M. Döpfner⁵,⁶, M. Romanos²

¹ Zentraleinstitut für Seelische Gesundheit Mannheim, Klinik für Psychiatrie und Psychotherapie des Kindes- und Jugendalters (KJP)
² Zentrum für Psychische Gesundheit, Klinik für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie, Universitätsklinikum Würzburg
³ Klinik für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie, Universitätsklinikum Gießen und Marburg
⁴ Klinik für Kinder- und Jugendpsychiatrie, Psychotherapie und Psychosomatik, LWL-Universitätsklinik Hamm der Ruhr-Universität Bochum
⁵ Ausbildungsinstitut für Kinder- und Jugendlichenpsychiatrie (AKiP) am Klinikum der Universität zu Köln
⁶ Medizinische Fakultät, Klinik und Poliklinik für Psychiatrie, Psychosomatik und Psychotherapie des Kindes- und Jugendalters, Universität zu Köln

Abstract: Objective: A recent Cochrane review published by O. J. Storebo and colleagues (2015) raised substantial doubts about the benefit from stimulant medication with methylphenidate in the treatment of childhood ADHD due to the overall poor quality of studies. The systematic review thus contradicts all previous reviews and meta-analyses. Method: We here detail various examples of errors, inconsistencies, and misinterpretations in the review which led to false results and inadequate conclusions. Results: We demonstrate that the study selection is flawed and undertaken without sufficient scientific justification resulting in an underestimation of effect sizes, which, furthermore, are inadmissibly clinically interpreted. The methodology of the assessment of bias and quality is not objective and cannot be substantiated by the data. Conclusions: Cochrane reviews lay claim to a high scientific quality and substantial relevance for evidence-based clinical decisions. The systematic review by Storebo and colleagues (2015) illustrates that, despite adhering to strict standards and high-quality protocols, even Cochrane works should be critically read and verified, sometimes with surprising results.

Keywords: Cochrane; review; meta-analysis; ADHD; methylphenidate

Despite their great therapeutic importance in the treatment of attention deficit hyperactivity disorder (ADHD), psychostimulants such as amphetamine and methylphenidate (MPH) have repeatedly been subject to public debate. A recent review published by a group of researchers led by O. J. Storebo (2015), which illuminates the efficacy and safety of MPH for the treatment of children and adolescents with ADHD, has now reignited this discussion. Indeed, in its edition from 25 November 2015 (p. 16), the German newspaper Südliche Zeitung (SZ) refers to this review in an article entitled “Dissipated Effect – Doubts About the Benefits of Ritalin and Similar Substances” [own translation], concluding that “The substances are taken by millions of children – but their benefits are limited, and side effects can strongly impair everyday life. Doctors therefore call for greater caution in prescribing Ritalin and co. for attention deficits and hyperactivity (ADHD) and for careful consideration of whether children and adolescents benefit

DOI 10.1024/1422-4917/a000433

© 2016 Hogrefe

Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie (2016), 44(4), e1–e8
from [the substances].” The SZ refers to an interview with one of the coauthors of the study: “Our expectations of the treatment are presumably greater than would be justified,” says the London-based child and adolescent psychiatrist Morris Zwi. “Although we can see hints of a benefit, the scientific evidence is of very poor quality.”

Negative reports in the media about psychostimulants are nowadays commonplace. However, the SZ news report is particularly noteworthy because Cochrane meta-analyses are generally perceived by experts and the public as representing established knowledge of the highest quality and with the greatest degree of evidence. The Cochrane collaboration is a global network of physicians, clinical researchers, methodologists, members of the healthcare professions, and patients. It advocates better health through improved possibilities for information. A central aim of the Cochrane collaboration is to create an evaluation of therapies oriented toward the principles of evidence-based medicine, to keep it up-to-date and to disseminate it. The systematic reviews published by the Cochrane collaboration claim to work in accordance with strict methodological rules in order to rule out biased results and systemic errors. Moreover, the first premise is to refrain from industrial financial support in order to ensure independence. Cochrane reviews therefore lay claim to a high scientific quality and substantial relevance for evidence-based clinical decisions.

The conclusion of the Cochrane review by Storebo and colleagues is nevertheless surprising, and in terms of the evaluation of the efficacy of MPH, it contradicts previous evidence. In most European countries, stimulants are approved as the first-line medication in children from the age of 6 years and adolescents as part of a comprehensive treatment program, inasmuch as low-threshold nondrug treatment measures such as psychoeducation, parent training, or behavioral therapy interventions (Eichelberger et al., in press; Mokros et al., 2015) alone prove to be insufficient. Psychostimulants are recommended by national and international guidelines for the treatment of ADHD in childhood and adolescence (Seixas et al., 2012). The most frequently prescribed medication in Europe is MPH.

The available evidence on this is extensive: Between 1962 and 1993 alone, more than 250 reviews and 3000 further single works on the effect of psychostimulants were published (Swanson, 1993). Between 1975 and 2007, the efficacy of these substances for the therapy of children and adolescents with ADHD was examined in over 180 randomized, double-blind, placebo-controlled studies, comprising over 12,000 children and adolescents, the results of which were published in peer-reviewed journals (Banaschewski & Döpfner, 2014; Walitza et al., 2016).

Changes due to treatment can be standardized by calculating effect sizes. The effect size is mostly defined as the difference in the mean changes of the experimental and control group, divided by the associated standard deviation at the end of treatment. An effect size of 1 indicates that the experimental and the control group differ from one another by one standard deviation (SD) regarding the measured variable. According to Cohen, effect sizes between 0.2 and 0.5 are small, those between 0.5 and 0.8 are medium, and those over 0.8 are large (Cohen, 1998). At the same time, however, effect sizes should not be equated with clinical relevance or significance, which we discuss in greater detail below.

Systematic meta-analyses and reviews, including the meta-analyses conducted by the National Institute of Clinical Excellence (NICE) (King et al., 2006), unanimously report effect sizes between 0.8–1 for the efficacy of MPH and amphetamine in terms of reducing the core symptoms of ADHD in children and adolescents (Bloch 2009; Charach, 2011; Charach, 2013; Faraone, 2002; Faraone, 2006; Faraone, 2009; Faraone, 2010; Hanwell, 2011; Kambeitz, 2014; King, 2006; Maia, 2014; Punja, 2013; Reichow, 2013; Schachter, 2001; Van der Oord, 2008); the strongest effects are found for attention, distractibility, impulsivity and social behavior in the classroom.

Storebo and colleagues (2015a), however, criticize that all 15 meta-analyses and reviews examining the available evidence on the efficacy and safety of MPH treatment in children and adolescents which they have identified so far, including the meta-analyses by NICE (King et al., 2006), showed considerable methodological weaknesses. In particular, they criticize that these works were not based on pre-published analysis protocols, and that the validity of the studies included in the works, and thus the possibility of distortion of results of the individual studies through various systematic errors (biases), were not, or not adequately, evaluated. They also claim that possible adverse drug reactions (ADRs) were not, or were only insufficiently, assessed. Moreover, the authors conclude that the fundamental benefit of medical treatment with MPH is not sufficiently substantiated. The Cochrane review was published by the authors as a summary in the British Medical Journal (Storebo et al., 2015b), where it came to generate highly contentious debate in the online comments section. In the following, we examine the methodology, results, and conclusions of the Cochrane review more closely with respect to its soundness and implications for scientific and clinical practice.

Why Does the Cochrane Review Find a Smaller Effect Size than Previous Meta-analyses?

The Cochrane review by Storebo et al. (2015a) comprises several meta-analyses, which included a total of 185 randomized, double-blind, placebo-controlled studies with
12,245 patients overall. It examined the efficacy and/or tolerability of MPH for the treatment of ADHD in children and adolescents. Thereof, 38 studies employed a parallel group design (n = 5111) and 147 studies (n = 7134) a crossover design. The efficacy of MPH was primarily examined based on teacher ratings. The authors calculated an average effect size of 0.77 (95% confidence interval 0.90-0.64) for the teacher-rated reduction of ADHD symptoms. However, this analysis included only 19 of the 185 studies, with a total of 1,698 patients – and exclusively those studies with parallel group or crossover designs for which the Cochrane authors had available data for the first phase (Biederman et al., 2003; Brown et al., 1985; Butter et al., 1983; Childress et al., 2009; Findling et al., 2006, 2008; Firestone et al., 1981; Ialongo et al., 1994; Jensen et al., 1999 (MTA Study); Kollins et al., 2006; Lehmkuhl et al., 2002; Moshe et al., 2012; Palumbo et al., 2008; Pliszka et al., 2000; Schachar et al., 1997; Taylor et al., 1987; Van der Meer et al., 1999; Wolraich et al., 2001). Moreover, the authors found an average effect size of 0.87 for the teacher-rated reduction of behavioral problems.

According to Cohen’s definition, an effect size of 0.77 is considered medium, though it only lies marginally below the previously described effect sizes of 0.8-1.0. Upon closer inspection, it becomes apparent that 4 of the 19 studies used to calculate this effect size did not have been entered into the analysis: In these studies, MPH was not tested against placebo, but rather a combined treatment with another intervention and MPH was tested against this intervention alone. These studies therefore resulted in clearly smaller effect sizes compared to the studies which tested MPH against placebo. In detail, these four studies tested clonidine + MPH vs. clonidine (Palumbo et al., 2008), parent training + MPH vs. parent training (Firestone et al., 1981), behavioral therapy + MPH vs. MPH (Brown et al., 1985; Jensen et al., 1999). In a further study of the Preschool ADHD Treatment Study (Kollins et al., 2006), the efficacy of MPH was examined in 3-6-year-old children; however, in this age group, MPH is not only an off-label medication, but is also generally less effective. According to our calculations, after excluding these five inappropriately included studies, the corrected effect size lies at 0.89 (95% confidence interval 0.79-1.01). According to Cohen’s definition, this corresponds to a large effect size and is in line with previous meta-analyses. Interestingly, it also roughly corresponds to the effect size which the Cochrane authors had calculated when they did not only include 19 studies, but rather all 75 studies in which teachers’ ratings on reduction of ADHD symptoms were available, i.e., after also including the crossover studies which only provided endpoint data (n = 6344; effect size: 0.91). Based on their hypothesis that crossover studies might lead to larger effect sizes due to a carry-over effect, Storebo and colleagues (2015a) had not taken these studies into account, even though they were unable to show any significant difference in the corresponding effect sizes (“we found similar treatment effects in the two groups and no significant subgroup differences,” p. 13; “we were not able to identify major differences when comparing parallel-group trials with cross-over trials”, p. 35).

Thus, it must be acknowledged that the selection of studies was in part manifestly flawed, and other relevant studies were excluded without sufficient reasoning. Despite shortcomings in the study selection, the calculated effect size lies only marginally below that found in previous meta-analyses. The inclusion of qualifying studies results in an effect size that corresponds to previous meta-analyses and systematic reviews.

How is the Clinical Relevance of the Effect Size Interpreted by the Authors of the Cochrane Review?

The effect size of 0.77 for the teacher-rated reduction of ADHD symptoms through MPH would, according to the Cochrane review, correspond to an average change of 9.6 points on the ADHD rating scale (DuPaul, 1991), which the authors interpret as a “moderate” improvement of symptoms. On the ADHD rating scale, the severity of symptoms is classified from 0 to 54 points – the authors of the Cochrane review erroneously state a range between 0 and 72 points here. This evaluation of the clinical significance by Storebo and colleagues (2015a) was based on a single study on the efficacy of atomoxetine by Zhang and colleagues (2005), with a total of 604 patients, in which the average change in the group means was related to the “minimal clinically relevant difference”. The change on the Clinical Global Severity Scale by one point corresponded here to a change of 6.6 points on the ADHD rating scale.

Storebo’s et al. (2015) evaluation of clinical relevance is dubious and must be rejected for several reasons. First, to determine the minimal clinically relevant difference, it is not methodologically sufficient to generalize from a single study to another class of substances - and to then draw far-reaching conclusions from this, with clinical recommendations. This surely does not correspond to the requirements of a Cochrane analysis. Moreover, the comparison of group means with the minimal clinically relevant difference, which only allows an estimation of the relevance of individual changes, is unsuitable and misleading for determining the relevance of group means, as was rightly criticized by the Quality Assurance Unit of the Institute for Quality and Efficiency in Health Care (Büchter & Thomas, 2015): “[R]egarding the effects on symptoms, measured...
on the ADHD rating scale or child health questionnaire, the authors compare mean differences to a particular, minimal clinical relevant difference for each of the scales. These measures, also known as 'minimal (clinically) important differences' (MCID or MID), are useful for judging individual change. However, it is misleading and of little help to patients and clinicians to judge mean group differences against an MID, because this ignores the variability in response.”

The interpretation of the clinical relevance of the (erroneously too small) calculated effect sizes in the Cochrane review is inadmissible. Derived from the published data, the benefit of treatment with MPH was called into doubt by the authors. This interpretation contradicts all international guidelines and is not based on sufficient evidence.

### Why is the Quality of All Studies Regarding the Efficacy of MPH Classified as Very Low?

The methodological quality of all 185 studies is called into question by the authors (Storebo et al., 2015a), who claim that 96.8% of the studies can be classified as being subject to a high risk of bias according to the Cochrane criteria, and that the further studies should also be seen as potentially biased. This gives rise to a considerable discrepancy between the quality assessments of the same studies by NICE (“moderate to high quality”) and by the Cochrane authors (“very low”). How can this be explained?

The bias assessment applied in the Cochrane review largely follows the relevant criteria and domains specified in the Cochrane handbook (Higgins 2011; domains: systematic error through inadequate randomization, insufficient blinding of study participants, study personnel and raters, group differences in the number of and reasons for drop-outs, biased assessment or selective reporting of results). For each domain and study, an evaluation of “low risk of bias,” “high risk of bias,” or “unclear risk of bias” is necessary. Storebo and colleagues (2015a) added an eighth dimension, “vested interest,” which is mentioned in the Cochrane handbook as optional. All studies funded by the pharmaceutical industry or whose study leaders declared potential conflicts of interest are therefore judged as being subject to a high risk of bias, irrespective of their scientific quality in all other aspects –even if no objections were found in all other assessment criteria. Moreover, the authors classified each study for which the assessability was deemed to be unclear in one of the eight domains as subject to a high risk of bias overall, and thus of insufficient quality.

Consequently, the studies for which no high risk of bias was found in any of the domains, but which, in the authors’ view, failed to mention necessary (detailed) information were also rated as being of poor quality in order to definitively rule out a systematic error in one or more areas. Regardless of whether this procedure can be seen as appropriate, the authors overlooked the fact that this information was actually provided in several works (e.g., Coghill et al., 2007).

For 184 of the 185 studies, the risk of bias for the various domains is presented in an overview (p. 27ff.). Thereof, 74 studies were judged not to have a high risk of bias in any of the domains; for a further 42 studies, a high risk of bias was only coded for the domain of vested interest; for a total of 68 studies, a high risk of bias was coded in at least one of the other areas; and in 29 of these, a high risk was also coded for the area of vested interest. For seven studies (not six, as was wrongly stated), the risk of bias was rated as low in all areas.

However, according to the authors, these seven studies are also possibly subject to a systematic error: They claim that the binding of raters is not assured, as ADRs occur more frequently in patients treated with MPH than those under placebo, which is easily identifiable and can lead to an overestimation of the efficacy and an underestimation of the risks of MPH (“It is likely that the trials initially judged to be at low risk of bias may, in fact, be trials at high risk of bias because methylphenidate gives rise to various prevalent and easily recognisable adverse events, which can lead to loss of blinding and hence can bias the ratings of symptoms, resulting in an overestimation of benefits and an underestimation of harms”; “it was possible for people in the trials to know which treatment the children were taking”). Contrary to these assumptions, ADRs such as reduced appetite, sleep disturbances, and difficulty falling asleep are not specifically associated with MPH, but are also seen under placebo conditions. They are not primarily observable in the school context. Should teachers assume in individual cases that children are being treated with MPH due to these phenomena, this also does not necessarily result in an overestimation of efficacy.

The overriding question then is whether the data provide hints that the efficacy and tolerability of MPH were systematically overestimated due to the studies classified by the authors as being subject to bias. In subsequent analyses by the authors, however, they found no indications that the studies judged as potentially biased differed from those for which this was not the case with regard to their results on efficacy (p. 25: “The estimated intervention effect did not vary according to risk of bias”; “No evidence suggested that the intervention effect varied according to risk of bias (low risk of bias versus high risk of bias”).

In summary, the Cochrane review produced a sweeping devaluation of the quality of decades of international research efforts and was undertaken with a degree of strin-
gency inappropriate even for the high requirements of Cochrane articles. On the other hand, the authors overlooked the presence of relevant information. Even using the strictest criteria, the reasoning behind the assessment of bias and quality is not sound. The fact that studies that otherwise met all criteria were ultimately also deemed to be of poor quality due to the potential unblinding danger arising from ADRs of MPH proves that the criteria applied by Storebo and colleagues (2015a) are inappropriate and cannot be satisfied. Finally, in their own analyses of the hypothetically assumed bias, the authors found no evidence for the presence of a relevant bias, but did not revise their evaluation accordingly.

How Can the Authors' Evaluation of the Domain Vested Interest Be Judged?

The funding of studies by industry is at times deemed to entail a large risk of bias. For instance, the authors of another Cochrane review (Lundh et al., 2012) came to the conclusion that the financial support of a study by manufacturers leads to clearly more positive results of an intervention, and thus appear to agree with Storebo and colleagues: “Our analysis suggests that industry-sponsored drug and device studies are more often favorable to the sponsor’s products than non-industry sponsored drug and device studies due to biases that cannot be explained by standard ‘Risk of bias’ assessment tools.” However, this principally applies to the interpretation of all study results. Nevertheless, the results of the Cochrane review by Lundh et al. (2012) also show that most meta-analyses do not find larger effect sizes in industry-funded studies.

Storebo and colleagues (2015a) erroneously report that 72 of the included studies were funded by the pharmaceutical industry. In actual fact, only 58 studies were sponsored by the pharmaceutical industry, some of which were approval studies (e.g., Coghill et al., 2013) and studies that were funded by companies that have no commercial interest in MPH as they produce alternative medications (e.g., Newcorn et al., 2008). In some studies only the study medication was provided by industry (e.g., Taylor et al., 2007). The approach of classifying industry-funded studies as generally subject to risk is therefore rightly criticized by Büchter and Thomas (2015a): “Instead of putting all trials with vested interests under general suspicion, we believe that it would be more reasonable to ask whether there are any reasons to believe that vested interests may have led to bias in each trial individually.” In response to this criticism, Storebo falsely claims that none of the studies funded by the pharmaceutical industry showed a low risk of bias in all other areas: “There were no trials with only the ‘vested interest bias’ domain assessed as ‘unclear risk of bias’ or ‘high risk of bias.’” Actually, however, at least 11 of the studies were classified as uncertain risk (Grzenko et al., 2012; Jacobi-Polisbouk et al., 2009; Tourette Syndrome Study Group, 2002; Riggs et al., 2011; Zeni et al., 2009) or as high risk (Arnold et al., 2004, Carlson et al., 2007; Kollins et al., 2006; Lehmkuhl et al., 2002; Stein et al., 2011; Waxmonsky et al., 2008) only in the domain of “vested interest.” Moreover, the authors could have statistically tested their hypothesis that studies funded by the pharmaceutical industry show larger effect sizes than studies without such funding, but they neglected to do so, or at least did not report this.

The general classification of industry-supported studies as subject to high risk was made in a sweeping manner in the Cochrane review, and the authors neglected to undertake a differentiated consideration, which might have led to a different outcome. In at least 11 studies, the classification “unclear” or “high” solely in the domain “vested interest bias” led them to be categorized as being of poor quality. The authors did not provide evidence for their assumption that industry-sponsored studies might generally be associated with more positive results in the studies.

What Conclusions Can Be Drawn?

According to Storebo and colleagues (2015a), the results of their meta-analyses suggest that MPH might improve teacher-rated ADHD symptoms, general behavior, and parent-rated quality of life in children and adolescents with ADHD. They claim, however, that the available evidence is of such poor quality that it is impossible to make any firm statements about the size of the effects. Moreover, they believe that it is generally unclear whether treatment with MPH is beneficial (p. 2): “[A]t the moment, the quality of the available evidence means that we cannot say for sure whether taking MPH will improve the lives of children and adolescents with ADHD.” The conclusions of Storebo and colleagues contradict all previous meta-analyses and reviews, including those by NICE, which provide evidence of a substantial efficacy of MPH in the treatment of ADHD.

The Cochrane review of the efficacy and tolerability of MPH treatment in children and adolescents with ADHD is marked by numerous inaccuracies, errors, and inconsistencies. Besides the aforementioned examples, there are also erroneous parameters for calculating the effect sizes of individual studies, such as for Jensen et al. (1999), Kolli ns et al. (2006), Moshe et al. (2012), Taylor et al. (1987), van der Meere et al. (1999). Further inconsistencies are found, for instance, in the bias classification for Konrad et
al. (2004) vs. Konrad et al. (2005), Ullmann (1986), Wallace (1994) vs. Wallander (1987), to name a few by way of example here. We have demonstrated that the study selection was flawed and undertaken without sufficient scientific justification. As a consequence, the effect sizes were calculated as too small and, moreover, were inadmissibly clinically interpreted. The methodology of the assessment of bias and quality is not objective and cannot be substantiated by the data. In particular, the aforementioned conclusion of the authors is grossly misleading. Finally, the conclusion that it is methodologically warranted to conduct so-called Nocebo studies, in which a substance which provokes solely undesired effects, claimed to be comparable with those of MPH, is given to the control participants instead of placebo, is extremely dubious from an ethical perspective.

Overall, it can be concluded that even Cochrane works should be critically read and verified. As recently demonstrated by Guy Goodwin (2015a), they are potentially misleading, for instance, if inadequate inclusion criteria are defined for studies, because dose-finding studies with too low dosages are used to assess efficacy, if studies that applied too high dosages are included to assess the frequency of ADRs, or if inappropriate criteria are applied to assess the validity of studies and available evidence is thus negated. The consequences of this become apparent in under-estimated efficacy, overestimated risks, and uncertainty on the part of clinicians and patients.

**Literature**


Storebø, O.J. (2015). Response to: Büchner, R.B., Thomas, S.: Re: Methylphenidate for attention-deficit/hyperactivity disorder in...
children and adolescents: Cochrane systematic review with meta-analyses and trial sequential analyses of randomised clinical trials, retrieved from http://www.bmj.com/content/351/bmj.j2503/rapid-responses.


Manfred Gerlach has received research grants pertaining to pharmacovigilance in children and adolescents from the German Federal Institute for Drugs and Medical Devices. He receives royalties from Springer Vienna for editing a German and English child and adolescent psychiatry textbook. He serves as the editor in chief of the journal ADHD Attention Deficit and Hyperactivity Disorders.

K. Becker has been involved in research/clinical trials with Eli Lilly and Shire, was on the Advisory Board of Eli Lilly/Germany, and was paid for public speaking by Eli Lilly and Shire.

Prof. Holtmann served in an advisory or consultancy role for Lilly, Shire and Medice, and received conference attendance support or was paid for public speaking by Bristol-Myers Squibb, Lilly, Medice, Neurocom, and Shire.

Dr. Doepfner served in an advisory or consultancy role for Lilly, Medice, Novartis, Shire and Viforpharma. He received conference attendance support, conference support or speaker’s fee by Lilly, Medice, Novartis and Shire. He is involved in clinical trials supported by Lilly and Vifor. He also receives honoraria as author of books and manuals on assessment instruments and on child behaviour therapy. He is also head of a School for Child and Adolescent Cognitive Behavior Therapy (AKIP) at the University of Cologne and he receives honoraria as consultant on Child Behavior Therapy for Kassenärztliche Bundesvereinigung (KBV).

Prof. Romanos served as consultant for the ADK Baden-Württemberg (German health insurance company). He declares no conflict of interest.

Prof. Dr. Med. Marcel Romanos
Zentrum für Psychische Gesundheit, Klinik für Kinder- und Jugendpsychiatrie, Psychiatrarie und Psychotherapie
Universitätsklinikum Würzburg
Füchsleinstr. 15
Deutschland
romanos@kip.uni-wuerzburg.de

Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie (2016), 44(4), e1–e8 © 2016 Hogrefe