Teaching Critical Appraisal of Articles on Psychopharmacology

Pavel Mohr, M.D., Ph.D.
Cyril Höschl, M.D., D.Sc.
Jan Volavka, M.D., Ph.D.

Objective: *Psychiatrists and other physicians sometimes read publications superficially, relying excessively on abstracts. The authors addressed this problem by teaching critical appraisal of individual articles.*

Method: *The authors developed a 23-item appraisal instrument to assess articles in the area of psychopharmacology. The results were collected with an electronic voting system. A discussion of each of the item followed; tutors shared their views and provided key ratings.*

Results: *Six publications were evaluated in the course of three workshops by a total of 58 trainees. Evaluation of the papers yielded varying results, reflecting variations of the participants' theoretical background as well as varied quality of the publications. The authors present detailed analysis of one paper as an illustrative example.*

Conclusion: *The discussion format and voting stimulated active participation of the trainees. Active involvement, facilitated by the structured assessment tool, followed by feedback with discussion, may enhance the learning process.*

Academic Psychiatry 2012; 36:114–117

There are numerous methodological problems in the published psychopharmacological literature. Some of these problems, including examples from published papers, are described elsewhere (1). Sponsorship bias and its effect on the wording of the report, particularly on the wording of abstracts, has been demonstrated (1). These problems are not always apparent to a superficial reader, particularly not to one who focuses only on the abstracts. In our experience, many practicing clinicians do not consider these problems important, thinking that these are "minor technical details." In this article, we will describe our effort to improve this situation by teaching systematic critical appraisal of individual publications.

Method

A series of workshops teaching critical analysis of published papers on psychopharmacology was conducted at three national psychiatric meetings in the Czech Republic and Slovakia in 2009 and 2010. The participants were psychiatric residents and other young clinical psychiatrists from both countries. The same three tutors (authors of this article) with experience in psychopharmacology were present at each teaching session to direct discussion and provide feedback to the participants. Each workshop analyzed two different papers that were sent to the participants before the event for individual study. One of the six papers selected by the tutors for analysis was industry-sponsored; five were independently funded.

The rating instrument was an appraisal sheet consisting of 23 items developed by the authors. The appraisal sheet was based, in part, on a similar instrument developed at the Center for Evidence-Based Medicine at the University of Toronto (*http://ktclearinghouse.ca/cebm/teaching/worksheets/therapy*); 22 items were questions exploring various aspects of the papers, with three possible respons-

Received July 15, 2010; revised August 23, 2010; accepted October 6, 2010. From the Clinical Division, Prague Psychiatric Center, Praha, Czech Republic; and the School of Medicine, New York University, New York. Send correspondence to Dr. Mohr; mohr@pcp.lf3.cuni.cz (e-mail).

es: Agree, Disagree, or Undecided. A fourth option ("Not Applicable") was added after the first workshop. One additional item was an unstructured question inquiring about further problems in the study and asking for comments. Full text of the appraisal sheet with an example of the results is shown in Table 1.

During the workshops, each paper was evaluated, one item at a time, by each individual participant, using the

TABLE 1. Appraisal Sheet With an Example of Results From a Single Evaluation

Item	Key Rating[a]			Participants' Rating (N = 27[b])			Agreement With the Key Rating (% of Correct Responses)
	Agree	Disagree	Undecided	Agree	Disagree	Undecided	
1. Was the hypothesis clearly stated?	1			24	0	2	92
2. Were subject selection criteria clear and appropriate for the hypothesis?	1			13	10	3	50
3. Was the number of subjects determined by power analysis?	1			22	3	1	85
4. Were all subjects who entered the trial accounted for at its conclusion?	1			24	0	2	92
5. Were intent-to-treat analyses performed?	1			24	1	1	92
6. Were per-protocol analyses performed?	1			21	4	1	81
7. Were the analyses dealing with dropouts adequate?	1			4	7	15	15
8. Aside from the experimental treatment, were the groups treated equally?			1	2	4	21	78
9. Were the drugs used administered at comparable doses?		1		5	19	3	70
10. Has compliance with medication been monitored?	1			17	3	7	63
11. Was the randomization adequate?	1			25	0	2	93
12. Was the blinding adequate (patients, raters, and clinicians)?		1		10	13	3	48
13. Were the assessment methods appropriate for testing the hypothesis?	1			23	0	4	85
14. Was the inter-rater reliability adequate?	1			20	0	7	74
15. Was the planned duration of trial adequate?	1			12	4	11	44
16. Did the statistical analyses test the hypothesis appropriately?	1			10	0	17	37
17. Did the statistical analyses test statistical significance?	1			23	2	1	88
18. Did the statistical analyses estimate the effect size?	1			19	3	5	70
19. If multiple test were performed, has a correction for multiplicity been made?			N/A	5	3	19	70[c]
20. Were analyses of adjunctive medications adequate?		1		0	26	1	96
21. Were ethical issues, including human subjects protection, appropriately addressed?	1			16	5	6	59
22. Do the authors' conclusions follow from their data?	1			13	8	6	48
23. Are there other problems with design or conduct of study?							

The data show cumulative votes for each item and agreement with the key rating.

[a] "Key rating" indicates tutors' responses.

[b] For Items 1–7 and 17, N = 26.

[c] "Undecided" was counted as "Not Applicable."

appraisal sheet. A direct-recording electronic voting system was used to collect and record the results. The text of the item was projected on a screen, and each individual endorsed one of the four answers (Agree, Disagree, Undecided, or Not Applicable), using buttons on a small recorder that was provided for each participant. After all the participants voted, the results in terms of frequencies for each of the four answers pertaining to the item being evaluated were instantly displayed on the screen and automatically recorded for future off-line analyses. At that point, a discussion of the particular item and its evaluation ensued. The tutors shared their views and provided feedback including the key ratings. The voting implied informed consent to participate in the workshop and was voluntary. Participants in the first workshop were supported by a pharmaceutical company (see Acknowledgment); the rest of the series was without any industry sponsorship.

Evaluation of one paper from a session with the highest number of trainees was selected as an illustrative example, and the following variables were assessed: 1) items with the highest agreement with the key rating (i.e., responses endorsed by the tutors), measured as a percentage of "correct" responses; 2) items with the highest rates of disagreement among the trainees, measured as a frequency of non-unanimous (discordant) ratings; and 3) items with the highest number of "Undecided" responses, where, in fact, the key rating indicated an "Agree" or "Disagree" response.

Results

Six papers were appraised in the course of three workshops. A total of 58 trainees participated. The methodological quality, as well as the clarity of the writing, varied greatly across the papers. The appraisal of one representative article was further analyzed (Table 1).

The results showed that:

1) The highest agreement with the key rating (above 90% of the "correct" responses) was achieved in items assessing adequacy of analysis of adjuvant medication (96%), adequacy of randomization (93%), clarity of the hypothesis statement (92%), whether all subjects were accounted for at the study conclusion (92%), and whether intent-to-treat analyses were performed (92%). By far the lowest agreement (15%) was found on the item evaluating adequacy of analyses dealing with drop-outs. Furthermore, poor agreement with the key rating (below 50% of the "correct" responses) was also observed on items appraising appropriateness of the statistical analyses testing the hypothesis (37%), planned duration of the trial (44%), adequacy of the blinding (48%), and whether the authors' conclusions stem from their data (48%).
2) The greatest disagreement among the raters, where the frequency of a single response (regardless of whether "correct" or "incorrect") was 50% or less, was found on items appraising subject selection criteria (Agree: 50%, Disagree: 38%, Undecided, 12%); blinding (38%, 50%, 12%, respectively); trial duration (44%, 15%, 41%, respectively), and whether the authors' conclusions followed from their own data (48%, 30%, 22%, respectively).
3) The highest rate of "Undecided" responses of the trainees was observed in items evaluating appropriateness of the statistical analyses testing the hypothesis (63%), adequacy of dealing with drop-outs (58%), and planned duration of the trial (41%).

Discussion

We have constructed an instrument for the appraisal of psychopharmacological publications and used it as a teaching tool in a series of sessions with psychiatric residents and other trainees. The instrument (appraisal sheet) highlights the principal methodological features that are important for the interpretation of the results of psychopharmacological trials.

The principal purpose of the exercise was to help the trainees to think critically about what they read. Using the systematic approach imposed by the checklist format of the appraisal sheet, they discovered for themselves various methodological problems in the publications that were scrutinized during the workshops. One of the principal functions of the tutors was to demonstrate how these problems affect the interpretation of the results. The discussion format and voting stimulated active participation of the trainees. We witnessed the result that active involvement, facilitated by the structured assessment tool and followed by immediate feedback and thought-provoking discussion, may enhance the learning process. Finally, our observations underscore the importance of teaching statistics and methodology.

Our feasibility study presenting the experience with this new teaching method has certain limitations. No meaningful statistical analysis of the results of the ratings was feasible because the numbers of papers and participants were too small. The results shown here thus serve as an illustrative example and do not allow for any generalization. Furthermore, we did not test the participants before and after the workshops to estimate the educational impact of their experience. These issues can be addressed in future research.

We thank Eli Lilly & Company for sponsoring the first of the three workshops. Drs. Mohr and Höschl were supported by the research projects CNS 1M0517 (from the Czech Ministry of Education) and MZ0PCP2005 (from the Czech Ministry of Health).

Dr. Mohr has received speaking honoraria and/or consulting fees and travel support from AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Janssen-Cilag, Krka, Servier, and Pfizer. Dr. Höschl has received speaking honoraria and/or consulting fees and travel support from Bristol-Myers Squibb, Eli Lilly, Janssen-Cilag, Gedeon Richter, Krka, Servier, and Lundbeck International Neuroscience Foundation. Dr. Volavka has received speaking honoraria and/or consulting fees and travel support from Eli Lilly and Merck.

Reference

1. Heres S, Davis J, Maino K, et al: 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 2006; 163:185–194