The Peer Review of “Outcome of Depression in Psychiatric Settings”

ALAN LEE and GORDON PARKER

To facilitate the introduction of this new section, designed to shed light on the peer review process, the Editor has used the review of a paper of which he is a co-author. The referees for this paper were originally anonymous, but they both agreed to be named for the purposes of this article. The original article was twice as long as the published version. The effect of the assessment should be readily apparent. Editorial decisions were undertaken by members of the Editorial Board other than the current Editor.

Alan Lee

This is an ambitious paper which compiles findings from all the reports of naturalistic follow-up studies of adults with depressive disorders that were published between 1970 and 1992. The authors acknowledge the vast degree of heterogeneity in this material, both in terms of methodology and follow-up periods (ranging from months to over 20 years), but attempt to turn this to their advantage. They do this by organising the findings into groups with comparable lengths of follow-up and taking median values for various outcomes among the reports in each group. In this way, general figures for the ‘average’ outcomes of depressive episodes are derived for percentage recovery, relapse, readmission, impairments of social functioning, chronicity, and suicidal behaviour. The separation of these concepts and the ‘results’ are presented in a clear, simple fashion which would help to popularise our increasing understanding of the long-term significance of episodes of depressive disorder.

The paper is lengthy and heavily referenced, and many of the potentially informative tables are extensive and very dense. As one might expect, making comparisons between and schematising reports of this kind is a difficult and often complex exercise, and some of this complexity makes heavy demands on the general reader. The overall messages of the paper, apart from those about suicidal behaviour, which appear plainly to be wrong, are not new and have been made in a much more tightly argued and concise fashion elsewhere. Those who are interested in finding out about the short- and long-term outcomes of depressive disorders might be better advised to read several of the more modern, methodologically sound, follow-up studies in their original form. As is perhaps not surprising given the simplifications and condensations necessary for such an overview, there is a sense in which in this paper the whole is much less than the sum of its parts and, indeed, less than several of the parts taken alone.

Part of the difficulty in evaluating this work follows from its ambiguous status. It is neither a definitive review nor a meta-analysis, and it runs the risk of falling between two stools. As a learned review, I am sure that the authors would agree that it does not stand comparison with the existing reviews, for example that by Jules Angst. In particular, there seems to be little systematic evaluation of the extremely variable methodological quality of the different studies, nor any rigorous exposition of the methodological issues involved.

Unlike in a meta-analysis, there is no apparent minimum standard of methodology (e.g. McMaster criteria) for entry into the field, so that weak findings are considered alongside strong ones and the median is taken willy-nilly. The median of X sheep and Y goats may be a sheep provided X is greater than Y, but if not the result will be a goat. Among these studies, those which have low follow-up rates, and those which have very variable follow-up lengths and fail to use an actuarial approach, are considered by most experts to be methodologically unsound. Any rates of recurrence, readmission and chronicity extracted from them are almost certainly meaningless. The risk of ending up with a median which is also meaningless is heightened by the authors’ practice of counting separately (up to five times in the case of the Iowa 500 series) multiple reports of the same methodologically unsound study. (To be fair to such studies, many of them did not set out to establish risks of recurrence and so on, and the data are often presented as incidental findings in a report which addresses a more answerable question.)

These problems are increasingly apparent as the paper moves towards longer-term studies, and the sections addressing these are rather weak. While the authors are aware of survival analysis and quote examples in shorter-term studies (e.g. Keller et al), there seems to be a limited understanding of the vital nature of such an actuarial approach to make sense of a longitudinal data set with considerable numbers of patients who may die or drop out of follow-up after widely varying intervals. Their failure to acknowledge this leads to meaningless comparisons of cumulative risk data with much simpler proportional data describing overall patterns of illness,
and this results in some bizarre inconsistencies (see below).

The weakest section of the paper is that addressing variations in outcome according to different subtypes. This is perhaps inevitable given the limited space remaining after assembling an account of the basic outcome data. The selection of findings on subtyping is patchy, and the discussion rarely does justice to the closely argued debates that characterise the literature in this area. The practice of dividing findings by follow-up period rather than subtype probably exacerbates this problem, as does the surprising subsequent confusion of long-term and short-term findings in the Discussion.

Two serious flaws in this exposition stand out. The first is the failure to consider whether the index episode is a first episode or not. One of the very few consistent, significant and useful predictive findings to emerge, particularly from longer-term studies, is that patients on their second or subsequent admissions have much greater risks of further readmission. The question of whether a follow-up of in-patients refers to a first admission or consecutive series is also, therefore, one of the most important variables to consider in evaluating and attaching weight to the outcome findings.

A second problem in the subtyping section relates to the endogenous–neurotic distinction. The authors claim that results here have been inconsistent, and that the terminology is outdated. In fact, the pattern of findings is entirely consistent: the short-term outcomes of endogenous depressives are better, but this group has a considerably higher risk of relapse, so that, in the longer term, global outcome is the same as that of neurotic depression, or possibly worse. The one robust finding which emerges from every analysis addressing this question is that endogenous or melancholic subtyping is a major predictor of higher rates of relapse. Since the effects of history and endogenous subtype have been shown in two independent studies to be additive and to have a very marked effect on readmission when taken together, it would be a major omission from a paper of this kind not to emphasise such a clinically valuable finding. In addition, the compelling results for endogenous or melancholic subtyping predicting better recovery and higher rates of relapse do contribute to the validation of the concepts; this picture has become clearer since Farmer & McGuffin's review, which is probably misleading now as regards the current status of this distinction.

The final area in which the conclusions of this report may be misleading concerns suicide and attempted suicide. This is a topical and vitally important element of the picture. Again, one of the major flaws in the review is apparent. Because little attention is given to the importance of an actuarial approach to the analysis of longitudinal data, many findings are reported which clearly underestimate suicide risk. It is also not clear that the authors have acknowledged that the rate of suicide will continue to rise as the period of follow-up increases, so that it is meaningless to aggregate findings from, say, a 7-year follow-up, a 16-year follow-up, and a mixed bag of subjects followed between 6 months and 20 years. The result of these shortcomings is that the conclusion that "about 1 in 20 commit suicide when monitored for 5 years or more" appears to be a major revision of Guze & Robins' widely accepted figure of a lifetime risk of 15% for depressive in-patients, but the new finding is probably unsound. The discrepancy needs careful discussion, as this is such an important topic and the results are likely to be widely quoted. If a revision is being proposed it probably warrants a short paper in itself, though it is likely that on more careful analysis the Guze & Robins figure will be seen to prevail. As regards suicide attempts, the treatment of these is clearly inadequate, there are too few studies to make a reliable estimate, and the suggested figures once again are likely to be too low by a factor of two or three. This section is so likely to be misleading and so unsound that it would be better omitted.

Conclusion

I believe that it would be unwise to publish this paper as it stands, given the limitations outlined above. I find it hard to advise how it can be easily modified to be suitable for the BJP, but nevertheless I think it would be worth an attempt, as the resulting paper could be an important one. There is a need for a much more penetrating critique of methodology, and an exposition of the pitfalls if an actuarial approach is not followed. Ideally, there would be some weighting of findings according to their methodological strengths, and an exclusion of some of the weaker studies, where the results are clearly meaningless. If some process of averaging data is followed (and the median is certainly an imaginative compromise, but not necessarily one resulting in a valuable synthesis), then results from the same series (e.g. the Iowa 500) should be considered only once. In order to avoid the resulting paper being impossibly long, it might be wiser to separate off the subtyping into a second paper and dropping the suicidal behaviour data altogether.

If the authors were prepared to attempt a revision then they might find the following additional detailed comments to be of some value.
(1) Standards for a follow-up study: criteria include prospectively defined cohort; standardised diagnoses; high follow-up rate; majority of series followed for the same period; survival analysis or equivalent to allow for deaths and attrition.

(2) Chronicity defined as "ill throughout follow-up" is a non-starter unless follow-up is of uniform length. Including these data tends also to muddy the picture and will be very hard for non-experts to follow. Might it be better to stick to 2 years' duration as a criterion?

(3) Why start the literature search in 1970? This excludes one or two important studies (e.g. Kay et al). Surely it is equally a mistake to include those papers which relate to the pre-treatment era (e.g. Opjordsmoen (published 1989, refers 1946–48), Stephens & McHugh (published 1991, refers 1913–40)).

(4) If the tables remain in a similar format, it would help to include follow-up length and number of patients and perhaps list in increasing lengths of follow-up.

(5) "Our terminology has a pragmatic descriptive purpose and does not signify operational criteria": is this not perhaps making a silk purse out of a sow's ear?

(6) More could be made of the relapse/recurrence distinction as this is now widely regarded as an important conceptual distinction with 'pragmatic' implications.

(7) The use of words such as "our findings", "our results", "our analysis" and the IMRAD layout of the paper suggest a spurious scientific status to the paper, which is plainly not a meta-analysis.

(8) The practice of picking on some (but not all) outliers for discussion in the results section is dubious and also gives emphasis to what is often later being dismissed.

(9) The reports of Kiloh et al and Andrews et al show some admission rates as higher than recurrence rates. This is clearly nonsense and illustrates the dangers of comparing sophisticated cumulative risk data from survival analyses with crude proportions showing different overall patterns (which can be very misleading due to deaths, attrition and other censored data).

(10) As an example of the problems of incomplete follow-up, 25% of Shobe & Brion's sample could not be traced, which limits and qualifies their findings of unusually good outcome.

(11) While not strictly within the brief of this paper, I think it would be valuable to mention the importance of personality as a determinant of outcome. All studies have shown neuroticism to predict poor outcome, possibly by leading to chronicity, and this effect appears to be independent of and additive to the other major predictors: endogenous/melancholic subtyping, history (i.e. already recurrent), and delusional subtyping.

References


Gordon Parker

The authors address a reasonably important issue in an ambitious way, but the paper is severely limited by its prolix style and relative failure to recognise a number of conceptual limitations.

At present, the paper reads like a literature review for a thesis, and is far too long for the B JP or, probably, for most other scientific journals.

The objectives are reasonable ones and the field is one worthy of attempt at synthesis. Nevertheless, there are intrinsic limitations to any attempt that seeks to generalise outcome statements about depressive disorders. The first concern (recognised by the authors but not resolved by the research community at this stage) is that depressive disorders are unlikely to be homogeneous, and that the outcome may vary considerably across differing depressive types. Secondly, the literature reviewed by the authors is generally very unclear about depressive types. Secondly, the literature reviewed by the authors is generally very unclear about remission, recurrence, relapse, and so on. The authors do not note recommendations by Frank et al that will allow for greater consensus in defining outcome markers, but the previous literature is flawed and limits precise statements. While both these points are conceded by the authors, a spurious precision is risked in their integrative analysis. Thus, they have laboured in an elephantine way but produced something less substantial. Additionally, I do not believe that they have given appropriate recognition to a number of other attempts at synthesising this literature. While they do make reference to Keller's survival analysis research, none of their references to Keller go beyond 1986, and he has produced a number of papers since then which address the same issue, and certainly with large numbers of subjects being considered in the analyses.

At a finer level of assessment, there are a few additional points. The writing is a little too staccato, often with a number of disparate points not always
closely linked. For instance, if we examine the Introduction, we see that the first paragraph addresses the prevalence of depression, the second paragraph the classification of depression, the third the outcome of depression – with no linking sentences explaining the relevance of these disparate points.

Additionally, I would quibble with a number of the statements which are used as sign-posts. In the first paragraph we are given lifetime prevalence data for unipolar major depression which are considerably lower than treated prevalence data for depression in the US, a potentially paradoxical issue not addressed at all.

The authors make confident interpretations which may not be shared by the reader. For instance, in the paragraph beginning at the bottom of page 3, we are told that the "most plausible explanation" is antidepressant drug being prescribed in an inadequate dose. An equally confident interpretation is that patients may cease to be compliant with medication after improvement.

The use of the term 'naturalistic' could be debated. At first pass, it would appear preferable to restrict the term to studies of 'untreated' depressive patients, and there certainly is a small list of such studies. The authors, however, use the term to include a number of situations, confounding interpretation.

I believe that the authors should be encouraged, firstly to reduce the paper by about two-thirds, synthesising the material to the nuts and bolts, and secondly, to concede a number of the theoretical and conceptual limitations to a far greater degree than they do in the present version.

The authors’ reply

We have rewritten this paper in line with the excellent comments made by your assessors and have cut the length by half, as you requested. Four of the tables and the two figures have been dropped.

I am enclosing three copies of our revision along with a commentary on the changes we have made.

The paper is much improved now and I look forward to hearing your response about the revision.

Alan Lee

We found the assessor’s response most helpful.

We have responded positively to most, if not all, of the numerous comments.

We have: acknowledged the importance of the actuarial approach; left out the sections on admission, suicide, and subtyping (but have discussed some issues relating to the last); tried (within the space constraints and the remit of the paper) to give a more penetrating critique; excluded weaker studies, weighted the findings, and counted studies only once at any follow-up.

We have also: (1) included standard of methodology; (2) made a distinction between persistent depression during follow-up and ‘chronicity’; (3) kept to 1970, for arbitrary reasons (but we take the point); (4) improved the table format as suggested; (5) omitted this phrase; (6) made more of the distinction between relapse and recurrence; (7) omitted suggestions implying spurious scientific status; (8) have not discussed outliers in the results when possible; (9) excluded this material; (10) excluded this article; (11) mentioned neuroticism.

Gordon Parker

Again, we found the assessor’s response most helpful.

We have virtually reanalysed and rewritten the paper, now emphasise the theoretical and conceptual pitfalls and limitations, and have updated the literature as suggested.

We have tried to improve the style, and we have eliminated the material that rightly was objected to.

(Paper reassessed by referees, and accepted)