Dear Sir,

A recent meta-analysis (Daya and Gunby, 1999) on recombinant versus urinary follicle stimulating hormone (rFSH versus uFSH) deserves methodological comment.

The *a priori* definition of eligibility criteria for studies to be included in a meta-analysis are central and must be systematic. It was up to the authors to decide whether their search would be extended to include all unpublished studies, but selective inclusion of some studies, performed with only one of the two recombinant compounds considered and, as it happens, amongst those investigations favouring this compound, is unacceptable. Ironically, the validity score of both unpublished studies had the highest rank order (Table II of the paper), making it difficult to understand the delay in publishing studies of such quality. In any case, this paradox confirms that scoring the quality of clinical trials for meta-analyses is a tricky job (Jüni *et al.*, 1999). Overall, this selection of two unpublished studies favouring follitropin alpha accounted for 40% of the patients receiving this drug; added to the fact that four of the six other studies on this compound came from simple abstracts, this means that 79% of the patients treated with follitropin alpha came from investigations which were not peer-reviewed (as compared to 13% in the follitropin beta series).

Table V of the paper suggests a comparison between follitropin alpha and follitropin beta. As everybody knows, this suggestion of a direct comparison between two agents which were not tested within a common investigation is illegitimate from a methodological standpoint. Secondly, the test of homogeneity made by the authors hid a striking heterogeneity between the series included, as the success rate ranged from 18 to 51% for rFSH (16-40% for uFSH) in those trials on follitropin alpha versus only 30 to 35% (27–28% for uFSH) in the studies on follitropin beta. In addition, the sole studies favouring uFSH as compared to rFSH came from the follitropin alpha’s series, one of them having been double-blind! Such difference in heterogeneity should have precluded any attempt to discriminate between both series – if not the meta-analysis itself. Thirdly, a simple inspection of data shows that the statistical significance favouring follitropin alpha and unduly put forward in Table V is due to the results of two studies which should not have been included in the meta-analysis, the first because it was one of the positive unpublished studies (Schats *et al.*), the second because it was not even randomised (Manassiev *et al.*, 1997). The weight of Schats *et al*’s investigation to suggest an average superiority of follitropin alpha over follitropin beta is all the more unacceptable since the results of this study were surprisingly poor, with a pregnancy rate on Gonal-F® of 25% (versus 16% on uFSH), i.e. 22% lower than the mean pregnancy rate on Puregon®, whereas the overall pregnancy rate was 30.9% in the follitropin beta series as compared to 27.9% in the follitropin alpha series. Actually, the only objective fact illustrated by Table V is a remarkable heterogeneity of the results obtained with follitropin alpha, which probably reflects a lack
of uniformity in the experimental design of the investigations performed with this compound (see Table III).

To sum up, the data collected by Daya and Gunby suggest that the results were on average slightly less favourable and impressively more fluctuating with follitropin alpha than with follitropin beta. As with every new methodological tool, meta-analysis carries a new potential of fallacies and this imposes considerable responsibility on the peer reviewers.

References

M. Girard*

1 Boulevard de la République, 78000 Versailles, France

*Dr Girard is an independent expert who was commissioned by Organon France to review the meta-analysis carried out by Daya and Gunby (1999). The decision to submit was his own, which Organon have asked to be made clear.

Dear Sir,

Thank you for the opportunity to respond to the several comments made by Dr Girard regarding our meta-analysis of recombinant FSH (rFSH) versus urinary FSH (uFSH).

We would like to start by reiterating that our search to identify all trials (whether published or not) that met our clearly specified inclusion criteria strategy was as comprehensive as possible. Consequently, we are surprised with Dr Girard’s assertion that there was selective inclusion of some trials. Since the publication of our study, we have identified several more unpublished trials which are currently being evaluated for possible inclusion in an updated meta-analysis; we are still awaiting responses from the investigators of two studies regarding our questions about their study methods and data and have contacted N.V. Organon for their assistance because follitropin beta was used in both studies. We would be grateful to both Dr Girard and N.V. Organon for drawing to our attention any trials they believe we have missed so that the necessary evaluation of eligibility can be performed.

The reliance on only peer-reviewed studies, as implied by Dr Girard (although it is unclear why he singled out for exclusion only one of the two unpublished studies, both of which had non-significant treatment effects) is problematic on two counts. First, for a variety of reasons, only a small proportion of studies presented at scientific meetings is submitted for publication. Thus, the potential for bias by not including the data from these studies is significant; publication bias of this type has been demonstrated to produce inaccurate estimates of the effect of treatment. Publication bias is a major threat to the validity and usefulness of a meta-analysis. The inclusion of data from unpublished trials is necessary to avoid this problem which may result in clinical practice being misdirected. Most meta-analysts are in agreement on this issue i.e. the search for evidence should be comprehensive and combine published data with data from what has been dubbed ‘fugitive literature’ or ‘grey literature’. This point is clearly evident from our study in which the two unpublished trials, which were completed very recently and are currently undergoing the peer-review process for publication, had the highest rank for methodological quality among all the trials. A difficulty arises in not being certain, despite extensive searching and consultation with the scientific community, that unpublished trials do not remain hidden. Furthermore, the willingness of investigators of located, unpublished trials to share their data can be a challenging obstacle to overcome. Second, the peer-review process is not always successful in ensuring that published results are valid, because it is neither structured nor consistent. Consequently, the published literature is replete with trials that have highly variable methodological quality. In contrast, the selection and evaluation process involved in a systematic review and meta-analysis is much more rigorous and reliable, because the criteria for selecting trials are specified a priori and at least two reviewers are required to independently evaluate the quality of the trials.

The issue of heterogeneity deserves clarification. Statistical heterogeneity is observed when the effect size from each trial varies, not only in magnitude (which is expected owing to sampling error) but also in direction. The problem is further magnified by conducting trials with small sample sizes. It may also be caused by known clinical or methodological differences among trials. In our study, the results were pooled using a fixed-effects model only after confirming that statistical heterogeneity was not present (i.e. the observed treatment effects in individual trials were not statistically significantly different from the overall pooled estimate of the treatment effect).

Dr Girard argues that the pregnancy rates for rFSH and uFSH among trials comparing follitropin alpha with uFSH were more variable than when follitropin beta was compared to uFSH. We would like to point out the fact that all three trials using follitropin beta (Out *et al.*, 1997a; Hoormans *et al.*, 1999) were multicentre trials comprising three, six and 18 centres, respectively. In the latter two of these trials (Out *et al.*, 1997a), information gleaned from the histograms in the publications showed remarkable variability in the pregnancy rates which ranged from 0% to 50% with rFSH and from 0% to approximately 46% with uFSH. Such wide variability, in part, may be the result of small sample sizes per participating centre, the range extending from three to 116 subjects.

There is no question that true randomization is a desirable method for generating comparison groups for clinical trials. However, the debate is still ongoing on whether quasi-randomization as was used in the Manassiev trial (Manassiev *et al.*, 1997) is also an appropriate method for this purpose. We believe that quasi-randomization methods do have a role but should be given a lower value in a validity scoring system as was done in our study. An issue of higher importance in reducing bias is that of concealing the allocation schedule.