February 13, 2003

Kathleen Reedy, M.S., R.D.H.
Health Scientist Administrator
Executive Secretary, Arthritis Advisory Committee
Dept. of Health and Human Services
Food and Drug Administration – HFD-21
Center for Drug Evaluation and Research
5630 Fishers Lane – Room 1039
Rockville, MD 20857

Re: Review of AVAVA

Dear Ms. Reedy:

I have been invited to provide an external expert review of an FDA memorandum dated November 7, 2002, co-authored by Drs. Renan Bonnel and David Graham and sent to Dr. Lee Simon, regarding ARAVA and its possible association with acute liver failure (ALF) and other serious hepatotoxicity. In my review, I will: 1) review and comment upon the methodology presented in this memorandum; 2) provide an interpretation of the results; and 3) offer some suggestions regarding alternative approaches for evaluating safety.

As Professor and Head of Biostatistics and Yale University School of Medicine, I have been involved in clinical studies research since 1977. My publication record includes development of new methods for clinical trials design, conduct, and analysis. Also, I have been co-author on numerous publications reporting results of clinical studies. My experience includes government, academic, and pharmaceutical-affiliated research activities, including both pre- and post-marketing activities. My full curriculum vitae are attached as an appendix.

A. Case Definition

The authors identified 102 cases from the FDA's Adverse Event Reporting System (AERS), and reviewed the events to quantify drug causality. The authors indicated intensive efforts were made to obtain followup from reporters and/or health care providers to aid in their assessment. This is an important task when conducting safety reviews. While the authors acknowledge the importance of obtaining followup, there was no evidence to suggest the nature and extent of these activities. Moreover, the memorandum concludes on page 17 that the case reports were of poor quality and many

reports lacked important data. "This combination imposed a significant handicap on risk assessment efforts."

Case definitions were provided on pages 4-5. For category 1 (non-fatal but severe and potentially life-threatening), it is not clear that "jaundice not requiring hospitalization" necessarily is a "non-fatal but severe and potentially life-threatening" event. Inclusion of such events, if not really severe, would inappropriately increase the total number of events. Moreover, this type of event does not seem equivalent in severity to their second criterion in this same category of liver injury requiring hospitalization. For category 2, the authors specify a list of liver-related events that must occur within 3 months of the development of liver-related signs or symptoms of jaundice in order to meet the criterion of "acute." The authors state that in some case reports, no timing was provided and classification was based on the narrative, suggesting a rapid time course. The number of such cases needs to be provided and results interpreted in light of the uncertainty associated with these data.

The causality assessments were categorized as: 1) probable/likely, 2) possible, and 3) unlikely. It is not stated whether the definitions were specified before the cases were reviewed, but it would be strongly preferred if the categories had been defined a priori. The definitions were vague, increasing the likelihood that subjectivity and bias would be introduced into the classification process. For example, "probable/likely" is an "hepatic event, including test abnormality, occurring in a plausible time relationship to drug administration." A "plausible time relationship" is vague; a precise time period should have been specified. The "event is unlikely attributed to concurrent disease or other drugs," is equally vague and subjective. A "possible" hepatic event included a lab test abnormality, occurring in a plausible time relationship to a drug. Again, no specification was provided as to what was "abnormal." For values outside normal limits, what did the laboratory value have to be before it was "abnormal" (e.g., 2X or 3X)? Was the test, if outside some limits, repeated? Finally, a "possible" hepatic event included subjects with "factors present that could plausibly have contributed to liver injury, but were not the most likely explanation for the adverse event." This vagueness invites misclassification and possible introduction of bias. In summary, the authors used liberal definitions of causality that tended to include cases, where others may have removed such cases.

Forty-eight cases were classified as "unlikely," and the remaining fifty-four cases were "possible" and "probable/likely" (page 5). Among the fifty-four cases, the authors stated, "The serious hepatic event was attributed to leflunomide and leflunomide was the primary suspect drug." This is an overinterpretation of seriousness and attribution based on the definitions above. Also, this result is misplaced in the methods section of their memorandum, since it precludes any analyses yet to be presented.

The authors point out further difficulties in ascertaining the nature and extent of liver adverse events, including but not limited to: 1) insufficient information was

provided to classify 35% of reports with respect to laboratory values, 2) liver biopsies were reported in only five patients, 3) data on the duration of ARAVA therapy prior to onset of liver injury was available in 41/54 patients, 4) specific information on timing for assessing temporality was not always provided and classification was only suggestive, and 5) forty patients received medication concomitantly that are labeled for hepatotoxicity. Concomitant drugs did "not appear to have been responsible for the acute hepatic event, or leflunomide was as likely to be the causative agent" (note this definition differs in a subtle way from "possible" above, where leflunomide had to be the more likely explanation).

In summary, every case report plays a key role in the calculations provided in the remainder of their memorandum. Thus, there is a need for clear, unambiguous definitions that would place all cases into appropriate categories for subsequent analyses and interpretations of results. There are substantive ambiguities in the definitions, assumptions that were made in classifying patients, and inadequate description of the steps taken to acquire additional case report information when information was lacking in order to reach a reliable conclusion. The nature and extent of requests made to clarify issues, and the success in acquiring additional information, would have been helpful in gauging the adequacy of the case report assessment procedures.

B. Epidemiologic Assessment

National Prescription Audit Plus data from IMS showed the number of US prescriptions of ARAVA from initial marketing in Fall, 1998 through May 2002. The authors used Tennessee Medicaid (TN) and United Health Care Group (UHG) data to determine persistency of drug use (a necessary prerequisite for their survival and hazard analyses). In their memorandum, Figure 1 shows the persistency of drug use from these two sources. The authors conclude the median duration of use was 4-5 months. Unfortunately, the authors do not define "persistency," nor do they indicate why they selected the TN and UHG databases. Other data sources should have been included (or at least mentioned) to provide an objective assessment of persistency, especially since the median drug use data reported here seems low.

I have been involved directly in research involving large databases such as TN and UHG, and persistency depends on numerous factors. Among these factors are switches and discontinuation of the drug. If a patient had a 2-3 week lapse between prescriptions, was this no longer "persistent"? Or, must discontinuation occur for some longer period of time before drug use is no longer "persistent"? If a subject switched to some other drug for a short period of time (e.g., 2 weeks), and then switches back to ARAVA, has persistency ended? What if a switch from ARAVA occurred, and the new drug was for long-term use taken for one month when an ALF endpoint was reached? Is the adverse event "probably caused" by ARAVA according to the methods in the memorandum? Or, was this assessment of relative causality made in the presence of both

drugs and their timing? Such clarifications are crucial in order to feel confident about categorization of case reports.

Moreover, persistency can be affected by issues unrelated to drug termination, including but not limited to: 1) formulary change, and 2) individuals switching to different health care plans. In summary, the persistency information in the memorandum requires further clarification in the presence of these real-life drug use practices and the peculiarities in which these data were collected and recorded. For, persistency plays a major role in the authors' survival and hazard function analyses. If persistency is longer than suggested by the authors, the risk estimates in their memorandum would be reduced, perhaps substantially.

For instance, assume that a patient on average takes ARAVA for 50%-100% longer than reported by the authors, a value consistent with other data sources. Then the number of person years-at-risk would be increased 50%-100%, to account for the longer duration of drug use. Thus, the numbers in Table 6 would be reduced. For example, for the "probable only ALF group," the values of 73-133 events per one million person-years (see Table 6) could be reduced to 37-67 events per one million person years (without changing the number of events). Moreover, the numbers in Table 6 are subject not only to "denominator" changes in person-years, but also to "numerator" changes in the number of case reports attributable to the drug. Reductions in the number of causally-linked case reports would reduce these values even further. In summary, the results in the memorandum are subject to wide variation. When one sees data that are subject to substantial variation as we see here, it is common practice to perform sensitivity analyses using various assumptions. Because this was not done here, the robustness of the results presented in Table 6 remains unexplored. It appears that the authors used "worst-case" estimates that may be prone to substantial bias.

C. Survival Analysis

Drs. Bonnel and Graham use the information described up to this point, to perform survival and related hazard analyses. The survival analysis and its implementation in this situation are non-standard.

In their survival analysis section, the authors draw from a number of different, noncomparable data sources (i.e., IMS, TN, UHG, and case reports from AERS) to obtain Table 7. This table gives the number needed to product one case report of ALF at 6 and 15 months. These results are obtained from a noncomparable array of different data sources, each with their own (unstated) strengths and limitations. A good deal of discussion (not provided) is needed to justify why data from disparate sources are combinable. Because the assumptions, coding, and methods used to collect data in each of these databases may not allow them to be combined for survival analysis, the survival analysis results should be viewed with caution.

Specifically, the authors use the persistency data, person-years of observation, and case reports (drawbacks of both noted above) to perform a survival analysis. The persistency data and person-years at risk are derived data, and not actual data as is the case in standard survival analysis. Thus, the objectivity and familiarity with standard survival analysis calculations is not present here. The consequences of using derived data include: 1) the estimates of variability and associated confidence intervals were not properly calculated in the memorandum, since these quantities ignored the extra variability associated with the derived nature of these data, 2) the survival and hazard estimates could vary substantially depending on selection of alternative data sources, and 3) use of derived data make specific estimates provided in the memorandum difficult to replicate independently (as I attempted briefly but unsuccessfully to do).

Their results should be made more transparent by exemplifying the calculations. Second, the persistency data, number of person-years at risk and true drug-related adverse events may not be representative. A more balanced analysis with statements of alternative assumptions and results would be desirable. A downward adjustment of their estimates of risk is likely, the magnitude of which could be large. Finally, the combinability for survival analysis is a major source of concern.

My final remark for this section is that one always would like to validate results from a complex modeling process with actual data. Tables 8 and 9 provide numbers that may be sufficiently small to check against the observed data in the TN and UHG databases. For example, Table 9 says that at 6 months the number of subjects needed to produce one case report of probable or possible ALF or other severe liver injury is 2,074. What do the data from the TN and UHG data show? The authors nowhere in this memorandum acknowledged validation as an issue. Comparison to known databases, either those presented here or others, is needed to reliably assess the safety data.

The section on hazard analysis is subject to the same issues as the survival analysis section above, and so I summarize my remarks here: 1) the need for transparency of calculations, 2) recognition that a number of different data sources (with unstated strengths and weaknesses) were required to perform calculations, and their combinability for survival analysis is a major source of concern, 3) the need for sensitivity analyses that evaluate other reasonable estimates of risk, persistency, and person-years of duration, 4) recognition of the need for model validation and, if possible, comparison to actual data, and 5) the width of the reported confidence intervals need to be broader to properly incorporate all variability.

D. Underreporting

The authors point out that underreporting is a serious concern, since serious adverse events are commonly not reported in the AERS. They refer to a number of well-known articles, many of which were published prior to the late 1990s. They update earlier tables in their memorandum with new tables that assume reporting of adverse event rates of 5%, 10%, and 25%. Thus, the respective increases in number of events are 20, 10, and 4-fold. The types of events discussed in this memorandum are dramatic and unexpected, so it is not clear that life-threatening ALF and related liver toxicity events are underreported to the magnitude suggested in the memorandum. In addition, there has been more recent and appropriate increased emphasis for reporting adverse events to the FDA. The magnitude of this problem, as it applies to this particular drug class or disease since 1998, also is not well documented. In summary, while an underreporting problem likely exists, it is sheer speculation what the rate is. Underreporting is simply another parameter in the modeling process that needs to be examined within a full sensitivity analysis.

The authors indicate (page 17) that ". . . the quality of cases can vary widely. Our overall assessment of leflunomide case reports of ALF and other severe liver injury was that they were generally of poor quality. There appeared to be little or no effort at followup for most of the cases reported. Associated with this, many case reports lacked important laboratory data as well as information relating to ultimate patient outcome. This combination imposed a significant handicap on risk assessment efforts." In summary, followup of case reports to adequately assess the nature and extent of liver-based adverse events seems unsatisfactory. The actual number of cases in the ARAVA safety evaluation is therefore subject to wide variation.

E. Additional Remarks

Because of ambiguities of case report data, and whether the primary analysis should be "probable plus possible" or "probable only," a sensitivity analysis was performed (see Table 12). Several analyses were performed, such as an exploratory analysis of the scores using a 25% percentile. There is no reason to select this particular cutoff; others are equally plausible and may lead to substantively different conclusions. Other analytic methods are presented, such as boxplots. But they add little to interpreting these data because the boxplots (Figure 6) summarize data from models that transform the raw vote data to a different scale; nothing new is achieved through their scoring system.

Inspection of Table 12 shows that there are three cases (of the sixteen) in which there is majority agreement by two distinct groups within the FDA. The authors state that "possible reports" were included in their analyses as linked to ARAVA, although

information provided in the case reports was incomplete (page 20). The authors conclude that excluding these events as drug-related would be biased, since additional information (if somehow gathered) would lead these reports to be upgraded in their level of certainty. This is speculation. My experience leads me to a different conclusion, where more information often can lead to excluding events, or to a reduction in the extent of causality. There is simply no way to make any firm conclusion about bias, and this argues for examining data from existing studies and/or new studies.

The authors evaluate methotrexate and liver injury (pages 20-23) to develop a risk-benefit analysis of ARAVA. However, these studies are from a different era, when adverse event reporting behavior was much different than today. Also, the nature and extent of collecting efficacy information probably differed greatly as well. So there is a lack of comparability between these older studies and the current ARAVA information. In addition, the methotrexate studies have very small sample sizes which preclude the identification of rare but significant adverse events. Another limitation is that "most sequential liver biopsy studies were based on relatively small numbers of patients' (page 23). Both of my issues are exemplified in the authors' remarks that "the best single study may be that of Buchbinder, et al., which appeared in 1993 and had 587 RA patients (page 22).

On page 24, the authors evaluate ARAVA data and indicate that the "last observation carried forward" overestimated the true response rate. Dr. Paul Leber and others at the DA years ago used this method, because it was a conservative approach to drug evaluation (i.e., it underestimated the drug's true efficacy). In the ARAVA setting, it would similarly underestimate the true response rate.

F. Recommendations and Summary

Additional efforts, beyond those presented in this one memorandum, are required in order to assess the safety profile of this drug. Additional studies need to be examined or performed. One type of study would be an observational cohort study, in which events are clearly defined and analyses performed. If at all possible, any events identified through a computer search would have the subject's chart available for review to confirm the diagnosis, the event temporality, etc. All concomitant and other relevant information would be assessed in order to eliminate the ambiguities as noted earlier in my report. Because very important adverse events found in large post-marketing experience are rare, it is not too great a burden to fully understand the nature and extent of all the treatment "twists and turns" and prognostic characteristics of each individual. Such information is crucial to determining the nature, extent, and degree of causality between the drug and the adverse event.

Another type of study would be a case-control study. A third type of study would be a properly conducted meta-analysis, which the FDA has relied upon in past reviews of drug safety. No matter what the study design, a well-designed study should have a

protocol that outlines all aspects of the study including hypotheses, endpoints, methods of case ascertainment and patient selection, etc.

In summary, the authors have attempted to address an important public health issue regarding drug safety. I raised questions regarding case ascertainment and assessment of the degree of drug causality. The resolution of how these cases are classified has a dramatic, multiple-fold effect on the stated results. Some of the case report deficiencies are intrinsically inherent in this type of investigation, however, and so these remarks are not intended as a criticism of the authors per se.

The authors have performed a useful attempt at addressing a difficult picture. However, the combining of disparate data sources required to perform survival and hazard analyses represent a novel analytic approach, which needs to be explored more fully before its reliability is answered. Also, the unstated underlying assumptions by the authors all seem intended to maximize the estimate of risk. I recommend that assumptions be explicitly stated and sensitivity analyses be performed that embrace the variability inherent in this process.

Because of these unresolved issues, these results represent a small piece of the safety puzzle for ARAVA. These data alone do not provide a sufficiently robust picture from which to draw full and reliable conclusions. Additional data and analyses from other sources are needed.

Sincerely,

Robert W. Makuch, Ph.D.
Professor and Head, Biostatistics
Yale University School of Medicine