Dr. Haseman was presented with the following background and questions:

ICCVAM has drafted performance standards for the traditional (radioactive) LLNA protocol that include a list of chemicals to be tested to demonstrate adequacy of an alternative method. There are 18 core chemicals, 13 of which are sensitizers based on LLNA results (i.e., they have an EC3 value that has been calculated). The metric for calculating accuracy is to test the 18 chemicals and get the 13 sensitizers within a range of EC3 values (i.e., 0.5 to 2x the historical mean EC3 value). The historical mean value is based on different numbers of studies, anywhere from 1 (5 chemicals only have 1 study and therefore the EC3 value generated in that one study is used as the "mean") to 49.

The Panel has asked that you consider what would be the likelihood that, if someone were to test these 13 sensitizers in the traditional LLNA, that they would get all of them correct using the 0.5-2x EC3 criteria?

Dr. Haseman's response:

It is important to understand that there are two performance standards mentioned in your E Mail, and they are NOT equivalent. The first is based on a statistical criterion of EC3 (or, ideally, log(EC3)) plus or minus X SD, where the value of X is chosen to reflect how stringent you want the performance standards to be. Selecting X=2 will result in approximately 5% of the studies being rejected as falling outside the acceptable range. If this results in "too many rejections", you can easily raise the value of X, perhaps to 3, as discussed below. The advantage of this performance standard is that you know fairly precisely the expected failure rate and can adjust the performance standard (i.e., change the value of X) to produce essentially any overall failure rate that is deemed acceptable.

As noted in my previous E Mail, the mean and SD for EC3 should be based on the log scale, not the original scale. That is, the acceptable range should be log(EC3) plus or minus X SD (the SD being based on the logged response), which makes it more comparable to the second performance standard discussed below. This is fairly important, since the distribution of EC3 responses on the original scale is somewhat skewed. Recognizing this skewness, the second performance standard discussed below assumes symmetry on a log scale (i.e., multipliers of 0.5 and 2.0 applied to the EC3) rather than on the original scale. The first performance standard should do this as well.

The second performance standard (0.5 to 2 x the EC3) has no obvious statistical justification that I can see. Using this standard, the likelihood of rejection is unknown and will vary from chemical to chemical, as discussed in more detail below. For some chemicals, it may result in no rejections at all. For other chemicals, the rejection rate may be as high as 40%. That is the reason I do not like this performance standard.

Please note that if X=3, and SD (on a log scale) = 0.231, then the two performance standards are identical, i.e., the first performance standard actually becomes 0.5 to 2.0 x EC3. If SD>0.231, then the second performance standard will result in more rejections than the standard based on log(EC3). If SD<0.231, the reverse is true. The difference is

that using the first performance standard, you know approximately the likelihood of rejection; for the second performance standard, you do not, since the underlying variability of the data is ignored.

Another matter that is unclear to me is what the consequences of "failure" may be. The panelist expressing concern implies that even a single failure will result in dire consequences for a lab. If such near-perfection is required (is it?), then I agree that choosing X=2 may result in too many rejections, but this can easily be remedied.

The panelist is correct, that if you use the plus or minus 2 SD criterion (applied to the logged EC3 response), then with 13 chemicals the likelihood is approximately 49% that one or more of the chemicals will "fail" the test (assuming independence). If this is deemed too stringent, then, as I said in my last E Mail, one possible solution is to impose wider acceptable limits. For example, if 2 SD was changed to 3 SD (i.e., X=3 rather than X=2), then for 13 chemicals the likelihood of an individual failure is only 0.26%, and the likelihood of at least one failure for 13 chemicals is

1-(.9974)13 or .03 or 3%.

Would the panelist find this 3% rejection rate for 13 chemicals acceptable? One can easily compute similar overall failure rates for different numbers of chemicals and different multipliers of the SD. It is also not possible to maintain the overall error rate at 5% as the number of chemicals increases, unless one is willing to change the multiplier of the SD after each chemical is tested, which is impractical

I have no idea what the corresponding failure rate is for the 0.5-2x EC3 criteria, since this criterion is not based on any formal statistical principles. As I indicated in my previous E Mails (and above), the chances of a failure would vary from chemical to chemical depending upon the inherent underlying variability in the estimation of the EC3. For those chemicals for which the EC3 can be very accurately estimated, the failure rate may be close to zero. For other more variable chemicals (e.g., sodium lauryl sulfate, based on the data you provided to me earlier) the failure rate for a single chemical may be as high as 40%. If there were 13 chemicals, each showing the variability of sodium lauryl sulfate, the overall likelihood of 1 or more rejections in 13 chemicals would be 99.9%.

In your E Mail your main question is "what would you consider to be the likelihood that, if someone were to test these 13 sensitizers in the traditional LLNA, that they would get all of them correct using the 0.5-2x EC3 criteria?" My answer based on the data you provided earlier (and the calculations presented above) would be somewhere between 0.1% and 99.9%, depending upon the underlying variability for the 13 chemicals. Not very helpful, is it?

In contrast, using a multiplier of the SD can control with reasonable accuracy the underlying failure rate, regardless of the underlying variability, and the multiplier can be chosen to be as conservative as you want. So if your question had been "what would you consider to be the likelihood that, if someone were to test these 13 sensitizers in the

traditional LLNA, that they would get all of them correct using the log(EC3) plus or minus X SD criterion", my answer would be: approximately 51% if X=2 and approximately 97% if X=3.

Moreover, for 100 chemicals (assuming independence and X=3), the probability is approximately 23% of a least one failure. That seems very low to me. Surely, a lab would not mind retesting one chemical in 100, and the chances would be 77% that even this would not be necessary.

For that matter, I am not sure that a lab would strongly object to retesting one chemical in 13 either (which has approximately a 50-50 chance using X=2), but I admit that this is a subjective judgment on my part, and I concede that I do not fully appreciate the (apparently) dire consequences of even a single failure for a lab.

The performance standard calculations using the SD multiplier assume independence from chemical to chemical, as noted by the panelist, but that is not an unreasonable assumption. That is, the likelihood of a failure for Chemical 2 is assumed to be independent of the likelihood of a failure for Chemical 1. It is difficult to envision how the results of tests carried out independently on different days could somehow be correlated. Of course, if a lab is truly incompetent, the chances of a failure for all chemicals may be greater than the corresponding probability for labs that do a good job, but that is OK and will weed out the weaker labs.

Another possible strategy that might reduce the overall number of failures is to recalibrate the acceptable range after each study. This may or may not be practical. If this is done, the multiplier of the SD would be unchanged, but the mean and SD would change, so the acceptable range would move slightly upwards or downwards as the additional data accumulate.

In my opinion, the important principle here is to use a performance standard with a known failure rate rather than one with an unknown failure rate. Thus, I strongly prefer the log(EC3) plus or minus X SD, rather than 0.5-2x EC3, where X can be selected to be as large as you like, depending upon how stringent you want the performance standards to be. A simple change such as selecting X=3 rather than X=2 as the multiplier of the SD would alleviate many of the concerns expressed by the panelist. If X=3 is viewed as an "over-correction", producing too few rejections, a smaller value of X could be selected.

For those chemicals that have only one study, calculation of an SD is impossible, so in this case, one might have no choice but to default to some arbitrary performance standard, such as the 0.5-2X EC3. Still, it should be understood that the failure rate for such chemicals is unknown. Thus, the concern that the Panel has expressed "that the requirement to get all 13 sensitizers correct and within a 0.5 to 2x range is too restrictive and may in fact be unattainable" is a valid concern, and one that I cannot address statistically, since this performance standard is somewhat arbitrary and the failure rate unknown and depends upon the specific chemicals tested. If I were forced to use this approach and asked to reduce the number of rejections, I would recommend widening the

acceptable range, perhaps to 0.3333 to 3 x EC3. This modified rule would still have an unknown failure rate, but it would produce fewer rejections than 0.5 to 2 x EC3.

Joe Haseman
2-22-08