May 2003

Estimation of the Underprediction Rates for the In Vivo Rabbit Dermal Irritation Assay

Prepared by Dr. Joseph Haseman
Consultant

I found the videoconference to be very helpful, and after hearing the various discussions and after consulting with Ray (who provided me with an updated database), I have fine-tuned the statistical analyses previously provided. Since certain of the panelists seemed confused by the false positive/false negative calculations, I will try to explain these in more detail.

APPROACH 1 (USING ALL THE DATA)

Here, my objective was to try to estimate for the universe of chemicals classified as "negative", "mild irritant" or "irritant", the distribution of animals showing a "score" below 1.5, between 1.5 and 2.3 or greater than 2.3 in each of the three classification categories noted above. If this distribution of probabilities could be derived, then it would be a straightforward matter to use these probabilities to calculate false positive and false negative rates based on a three animal screen. In this three animal screen, only ten outcomes are possible, and these are given below, together with the assigned classification for each outcome.

	Score		Classification
<1.5	1.5<2.3	>2.3	
3	0	0	Negative
2	1	0	Negative
2	0	1	Negative
1	1	1	Mild irritant
1	2	0	Mild irritant
0	3	0	Mild irritant
0	2	1	Mild irritant
1	0	2	Irritant
0	1	2	Irritant
0	0	3	Irritant

Since we (at least I) do not know the true classification of the chemicals independently of the data, I have made the simplifying assumption that the test results are "correct" for singly tested chemicals, and when disagreements occur among multiply tested chemicals, the "majority" call prevails; if there is a tie, the stronger call prevails. I concede that this approach may produce some under-estimation of the true underlying false positive and false negative rates, since there is a certain circularity of reasoning in assuming that the procedure is 100% accurate and then estimating the false negative and false positive rates. However, without multiple testing of chemicals (or independent verification of the accuracy of the procedure), I see no way out of this problem other than using only the multiply tested chemicals (see Approach 2 below). Differences between this revised analysis and the previous one are noted below:

(1) The revised list of chemicals provided to me excluded several chemicals classified as "corrosives", since Ray indicated that these would not have been subjected to irritancy testing.

(2) Test results for erythema and oedema were summarized SEPARATELY for each experiment, and the most severe outcome pattern of the two was used as the basis of classification for a given experiment. This differed from the approach used previously, in which the most severe of the two scores for each ANIMAL was used as the basic of classification, which allowed theoretically for the classification of a given chemical to be based on a mix of erythema and oedema. This particular refinement had essentially no impact on the chemical classification, but it did have some impact on the distribution of scores for a few chemicals.

(3) For those experiments utilizing more than three animals, the classification of the chemical was based on a random sample of three of the animals tested. In general, this had little impact on the analysis, but in some cases, it affected the classification of the chemical.

For example, there were quite a few 4-animal experiments in which the outcomes for the 3 categories (Score <1.5; Score <2.3 but >1.5; Score>2.3) were 0-2-2 respectively. Previously, this chemical was classified as an irritant, since there were two scores exceeding 2.3. In the revised analysis, this chemical was classified as an irritant half the time and as a mild irritant half the time, since a random sample of three animals from the four would result in a response of 0-1-2 (irritant) half the time and 0-2-1 (mild irritant) half of the time.

(4) To simplify the statistical analysis, I elected not to calculate the false positive and false negative rates for what I defined previously as Variation 1 of each approach. It had little impact (involving only three chemicals) and was just too time-consuming.

(5) However, I replaced this variation with another one, which involved how to weight each chemical's contribution to the estimate of the underlying distribution of test scores. For (new) Variation 1, each chemical was given equal weight. In Variation 2, the chemicals tested multiple times or with more than three animals were "weighted" proportionally to the number of animals used. There are advantages and disadvantages to both approaches that could be discussed in more detail if deemed necessary. One issue is whether the multiply-tested chemicals are a true random sample of chemicals or are they disproportionately represented by chemicals that are difficult to classify accurately/consistently.

From a practical point of view, Variation 2 produced slightly higher false positive and false negative rates, suggesting that those chemicals multiply tested tended on average to give slightly more variable results than those singly tested. Results are given below

APPROACH 1 VARIATION 1, CHEMICALS GIVEN EQUAL WEIGHT

		True class of chemical		
		Negative	Mild irritant	Irritant
Our decision	Negative	99.1%	3.8%	0.2%
as to class	Mild Irritant	0.9%	95.4%	8.8%
of chemical	Irritant	<0.1%	0.8%	91.0%
Estimated probability of				
An animal scoring < 1.5		94.40%	11.74%	2.60%
An animal scoring > 1.5 & < 2.3		5.07%	83.06%	15.91%
An animal scoring > 2.3		0.53%	5.20%	81.49%

APPROACH 1 VARIATION 2, CHEMICALS WEIGHTED BY NUMBER OF ANIMALS TESTED

		True class of chemical		
		Negative	Mild irritant	Irritant
Our decision	Negative	97.8%	4.8%	0.4%
as to class	Mild Irritant	2.2%	94.0%	15.8%
of chemical	Irritant	<0.1%	1.2%	83.8%
Estimated probability of				
An animal scoring < 1.5		91.24%	13.21%	3.43%
An animal scoring > 1.5 & < 2.3		8.37%	80.19%	22.06%
An animal scoring > 2.3		0.39%	6.60%	74.51%

APPROACH 2 (USING ONLY THE CHEMICALS MULTIPLY TESTED)

Using Approach 2 the objective is not to estimate the distribution of the possible outcomes, but rather to simply calculate the agreement/disagreement among multiple tests, assuming the most frequent outcome is "correct" (stronger outcome in case of ties). Here again, the results of the previous analysis were modified to a 3-animal test. Results are summarized below, together with a summary of the test results for the 23 multiply-tested chemicals.

		True class of chemical		
		Negative	Mild irritant	Irritant
Our decision	Negative	100.0%	6.9%	0.0%
as to class	Mild Irritant	0.0%	87.9%	27.8%
of chemical	Irritant	0.0%	5.2%	72.2%

Observed outcomes	Frequency	Classification
Neg/Neg	8	Negative
Neg/xxx	2	xxx
Mild/Mild	3	Mild irritant
Mild/xxx	1	Mild irritant
Mild/Mild/Mild	3	Mild irritant
Mild/Mild/yyy	1	Mild irritant
Mild/Mild/zzz	1	Mild irritant
Mild/Mild/Mild/yyy	1	Mild irritant
Irr/yyy	1	Irritant
Irr/irr/yyy	1	Irritant
Irr/Irr/Mild/yyy	1	Irritant

xxx: negative half the time; mild irritant half the time
yyy: mild irritant half the time; irritant half the time
zzz: negative half the time; irritant half the time

I conclude that the various refinements had minimal effect on the overall false positive and false negative rates. Approach 2 has slightly higher false negative/false positive rates than Approach 1, as expected, and in my opinion is probably closer to producing the "true" values, although more multiple-experiment chemicals are needed to be confident in this conclusion. For these data, irritants can easily be mistaken for mild irritants, but the likelihood of a mild irritant being a false negative appears to be less than 10% and the likelihood that an irritant will be falsely labeled negative seems close to zero.

Joe Haseman