Comment on

Houde, D.; Berkowitz, S. A.; Engen, J. R.,
The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies.
J. Pharm. Sci. 2011, 100, 2071-2086.

David D. Weis
Department of Chemistry, The University of Kansas, 1567 Irving Hill Road, Lawrence, KS 66045
dweis@ku.edu

In 2011, Houde, Berkowitz, and Engen published an important foundational work on the methodology of HX-MS for differential measurements in comparability contexts. Essentially, this work was perhaps the first to rigorously address the critical question 'what is the smallest difference in an HX-MS measurement that we should classify as significant?' At the end of the work, Houde et al. conclude that if the difference is greater than 0.5 Da in triplicate measurements, it should be considered significant.

In the course of a critical revisiting of this paper in preparation for a forthcoming manuscript, I have discovered that, while the overall premise of this work is sound, there are some errors in the statistical implementation. Overall, it appears that these errors will lead to a substantial overestimate in the magnitude of the significance limits when the methods described in the paper are followed. This Commentary is intended to provide guidance on how to avoid these errors.

It appears, based on my reading of the text, that three different statistical errors were introduced. Two of these errors will lead to overestimating the significance limit and one leads to an underestimate. Overall, it appears that the effect is that the threshold for statistically significant differences is much more stringent than was intended with the use of a purportedly 98% confidence interval. Overestimating the limit for significance has the apparent benefit of decreasing false positives (i.e., type I errors), but the undesirable side-effect of a loss of power due to an increase in the number of false negatives (i.e., type II errors). Thus, there is an increased risk of missing significant differences that appear to be 'too small' based on an overestimated significance limit.

Rather than just proclaiming where I think the errors are, I have gone into some detail to explain these statistical issues at a level that should be approachable by anyone who is familiar with mean, standard deviation, and Student's *t*-test. Recommendations are provided at the end.

Notation

Since the issues addressed in this commentary hinge on technical matters around the statistical treatment of data, I have recast the equations in notation that is commonly employed in the field of statistics. Houde et al. define the HX difference between the reference (subscript "ref") and experiment (subscript "exp") as an "array of differences", D

$$D\left(\Delta M_{i,t}\right) = S_{\text{ref}}\left(M_{i,t}\right) - S_{\text{exp}}\left(M_{i,t}\right) \tag{1}$$

where $S_n\left(M_{i,t}\right)$ is an array of mass increases for peptides indexed by their location in the amino acid sequence, i, at HX labeling time t, for data set n where n is the label for the reference or the experiment. Since the issues raised here do not concern the identity of the peptide or the HX labeling time I have dispensed with labels and recast the equation as

$$F = X - Y \tag{2}$$

where X represents the HX data from reference sample and Y represents the HX data from the experiment sample.

Determination of the standard deviation of replicate differential HX measurements

The first matter that arises is how the mean difference was determined, and on this point, it seems that the paper is unclear. Technical replicates were obtained so that we can write $X = \left(X_1, X_2, \ldots, X_n\right)$ and $Y = \left(Y_1, Y_2, \ldots, Y_n\right)$ where each subscripted symbol represents a single experimental determination of the amount of deuteration in a particular peptide at a specific HX labeling time. In the work of Houde et al., $n = 3$. According to the Supporting Information (pp. 4-5), F was "obtained from the average of three separate H/DX-MS measurements conducted on the same sample." Hence we have

$$\bar{F} = \bar{X} - \bar{Y} \tag{3}$$

where the overbar represents the arithmetic mean. It is trivial to show that $\bar{F} = \overline{X - Y}$:

$$\bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right) = \overline{X - Y}. \tag{4}$$

Plainly speaking, the differences of the means of X and Y is equal to the mean of the differences. The first problem that arises in this work, however, comes when one begins to reckon with the variance, or equivalently, the standard deviation. The exact definition of the standard deviation was not stated in the paper, but most experimental scientists would expect the sample standard deviation with the Bessel correction[1]

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}. \tag{5}$$

Since the square root notation is cumbersome, in many contexts it is preferable to refer to the sample variance, s^2. It is also worth noting here that the sample standard variance is what is known as a point estimator of the true variance, $\hat{\sigma}^2$:[1,2]

$$\hat{\sigma}^2 = s^2 \tag{6}$$

that is based on a limited number of replicated measurements. Before we form the standard deviation expression for F, it is useful to consider the error propagation in F:[3]

$$\sigma_F^2 = \left(\frac{\partial F}{\partial X}\right)_Y^2 \sigma_X^2 + \left(\frac{\partial F}{\partial Y}\right)_X^2 \sigma_Y^2 + 2\operatorname{cov}\left(X,Y\right) \tag{7}$$

which in the present case becomes

$$\sigma_F^2 = \sigma_X^2 + \sigma_Y^2 - 2\operatorname{cov}\left(X,Y\right) \tag{8}$$

where the function cov is the covariance of X and Y:

$$\operatorname{cov}\left(X,Y\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \tag{9}$$

expressed here with the $n-1$ Bessel correction. In the case of random, uncorrelated errors, which it is reasonable to expect in HX-MS measurements, the covariance would be zero, and replacing the variances with their point estimators gives the "common sense" expression for the propagated error in F:

$$\hat{\sigma}_F^2 = \hat{\sigma}_X^2 + \hat{\sigma}_Y^2 = s_X^2 + s_Y^2. \tag{10}$$

Importantly, equation (10) is valid if and only if X and Y are ***uncorrelated*** variables. In the Supporting Information, Houde et al. stated that they "determined SD values for all calculated mean $D\left(\Delta M_{i,t}\right)$ data points" (p. 4) which I interpret as 'the sample standard deviation was determined from the HX ***differences***.' This is a point where there is a lack of clarity in the paper, there are two ways to go about this, and looking initially at equation (4), it would appear that the difference is unimportant. Based on my reading of the text, it seems that the Houde et al., estimated their variance, denoted here as ε_F^2, using the mean of ***paired*** differences:

$$\varepsilon_F^2 = \frac{\sum_{i=1}^{n}(F_i - \bar{F})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - Y_i - \bar{X} + \bar{Y})^2}{n-1} = s_X^2 + s_Y^2 - 2\operatorname{cov}\left(X,Y\right). \tag{11}$$

This approach represents computing the differences from three paired measurements, i.e., $F_i = X_i - Y_i$, taking their mean value ($\bar{F}$), and computing a sample standard deviation. However, based on (10) the propagated error would be expected to be

$$\hat{\sigma}_F^2 = s_X^2 + s_Y^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}. \tag{12}$$

It is clear by inspection that equations (11) and (12) are not equal, i.e., $\varepsilon_F^2 \neq \hat{\sigma}_F^2$. Thus ε_F^2, computed as suggested by Houde *et al.*, is a biased estimator[1,2] of the variance in the HX difference and should not be used. The problem arises from the artificial correlation introduced when X_i and Y_i are paired to calculate the mean difference, $\overline{X-Y}$, rather than the difference of the means, $\overline{X}-\overline{Y}$. Again, the HX difference, $\overline{F}$, is not affected, as shown by equation (4), but the error estimate based on equation (11) is biased by covariance.

A simple numerical example can also serve to illustrate this issue. Consider $X=(1,2,3)$, $Y=(1,2,3)$, $Z=(3,1,2)$ where the only difference between Y and Z is that the order of the results has changed. It can be shown that $\overline{X}=\overline{Y}=\overline{Z}=2$ and thus that $\overline{X}-\overline{Y}=0$ and $\overline{X}-\overline{Z}=0$. Also, $s_X=s_Y=s_Z=1$. The propagated error, based on equation (10), is $\hat{\sigma}_{X-Y}^2=\hat{\sigma}_{X-Z}^2=1^2+1^2=2$. However, in the case of $\overline{X}-\overline{Y}$, computing the propagated error by (11), in other words, the mean of the differences, gives

$$\varepsilon_{\overline{X-Y}}^2=\frac{\left((1-1)-(0-0)\right)^2+\left((2-2)-(0-0)\right)^2+\left((3-3)-(0-0)\right)^2}{3-1}=0.$$

While for $\overline{X}-\overline{Z}$, the result is

$$\varepsilon_{\overline{X-Z}}^2=\frac{\left((1-3)-(0-0)\right)^2+\left((2-1)-(0-0)\right)^2+\left((3-2)-(0-0)\right)^2}{3-1}=3.$$

Thus we can see that the error propagated by equation (11) produces inconsistent results that depend on the arbitrary ordering of independent observations. Inclusion of the sample covariance, $\operatorname{cov}(X,Y)=1$ and $\operatorname{cov}(X,Z)=-\frac{1}{2}$, (i.e., equation 13) corrects for the artificial correlation, resulting in the "common sense" error of $\hat{\sigma}^2=2$ under both scenarios $\overline{X}-\overline{Y}$ and $\overline{X}-\overline{Z}$. Thus the important imperative here is that for differential HX-MS measurements, the two conditions must be treated as independent, uncorrelated statistical entities or the error analysis must include the covariance. In other words, propagate error based on the difference between the means rather than the mean difference. In some cases, neglecting covariance in paired data will lead to an overestimate of $\hat{\sigma}^2$ and in other cases it will be an under-estimate, depending on whether the pairing results in a positive or negative covariance. In the limit of very large numbers of technical replicate measurements of uncorrelated variables X and Y, the covariance will approach zero, but in small samples, such as triplicate measurements, there will usually be some amount of covariance. By appropriately averaging a large collection of independent determinations, the averaging would cancel out the covariance since the covariance is equally likely to be positive or negative. However, as described in the next section, the method used to determine the average standard deviation must be chosen appropriately.

Estimating the standard deviation from a population of measurements

Houde et al. proposed using the entire collection of sample standard deviations obtained from all of the replicate differential measurements to improve the reliability of the estimate. This is highly commendable because estimates of error based on triplicate measurements are notoriously unreliable. Setting aside the bias introduced by omission of covariance as discussed in the preceding section, there is an additional problem here, however. The collective standard deviation obtained was the "simple average of all these individual experimentally determined SD values" (SI, p. 4). Here, I assume that by "simple average" the authors are using the arithmetic mean. Taking the arithmetic mean of standard deviations is not a statistically accepted method to estimate the population standard deviation, even if the experimental error were estimated in replicates using either equation (5) or equation (10). Instead, a pooled estimate of the standard deviation, s_p, should be used:[2]

$$\hat{\sigma} = s_p = \sqrt{\frac{\sum_N (n_i - 1) s_i^2}{\sum_N (n_i - 1)}} \tag{13}$$

where s_i is a standard deviation obtained from replicate measurements, n_i is the number of observations, three in in the work of Houde et al., and $N = 670$, based on 67 peptides observed at five HX labeling times in two conditions as suggested by equation (4) of the paper. When all n_i are equal, equation (13) reduces to the root-mean-square (RMS) of the standard deviations:

$$\hat{\sigma} = \sqrt{\frac{\sum_N s_i^2}{N}}. \tag{14}$$

The well-known arithmetic mean RMS inequality states that:

$$\frac{\sum_{i=1}^{n} X_i}{n} < \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}} \tag{15}$$

for $X_i > 0$. This inequality indicates that using the arithmetic mean will lead to an underestimate of the experimental error. Thus, the mean value of 0.14 Da reported by Houde et al. underestimates $\hat{\sigma}$, the error of the differential measurements. The pooled estimate could be based on X and Y separately or the differences of the averages, $\overline{X} - \overline{Y}$, however, as suggested in the following section, working with differences leads to confusion in the application of Student's *t*-test that Houde et al. used to establish a 98% confidence interval.

Setting a threshold for a statistically significant difference using Student's t-test

Following the averaging of standard deviations, Houde et al. propose a form of the Student's *t*-test to set a threshold for a statistically significant HX difference. In other words, to determine if

$\left|\bar{X}-\bar{Y}\right|$ is large enough to be statistically significant. Although the approach was not described in these terms, Houde et al. used a two sample *t*-test for comparison of means assuming equal variance. The null hypothesis in this test is that the two samples, X and Y, are drawn from the same population. Loosely speaking, the null hypothesis is rejected when $\left|\bar{X}-\bar{Y}\right|$ exceeds a critical value, the conclusion being that the difference between X and Y is statistically significant. To perform a *t*-test, one chooses a desired $(1-\alpha)\times 100\%$ confidence level. The critical threshold for significance is then defined as

$$\left|\bar{X}-\bar{Y}\right| > t_{\frac{\alpha}{2}:df}\, s_p \sqrt{\frac{1}{n_X}+\frac{1}{n_Y}} \tag{16}$$

where $t_{\frac{\alpha}{2}:df}$ is the Student's *t* value based on $\alpha/2$ and the degrees of freedom, df. Houde et al. selected $\alpha = 0.02$ to obtain a 98% confidence level. In the implementation by Houde et al., the following equation was formed

$$\left|\bar{F}\right| > t_{0.01:2}\,\bar{\varepsilon}_F \sqrt{\frac{1}{3}} \tag{17}$$

since $n_X = n_Y = 3$ and the error estimate was based on the difference. The preceding two sections have highlighted difficulties associated with using the arithmetic mean experimental error $\bar{\varepsilon}_F$. Yet, even putting aside the problems associated with the error estimate, $\bar{\varepsilon}_F$, there is an additional problem with the threshold in equation (17): the degrees of freedom is incorrect. In a two-sample *t*-test with equal variance

$$df = n_X + n_Y - 2 \tag{18}$$

and thus $df = 4$ for the difference between means from triplicate measurements. The difference here is substantial: $t_{0.01:2} = 6.965$ while $t_{0.01:4} = 3.747$ leading to an almost two-fold overestimate of the magnitude of the significance limit.

Recommendations

The overall premise of the work by Houde et al., is sound: use a large population of replicate measurements to establish a significance threshold for differential HX measurements. However three errors in the implementation will lead to an erroneous overestimate of that significance limit. To avoid these errors, three simple corrections are needed:

1. Compute the individual standard deviations based on the sample means rather than the paired differences.
2. Use the pooled standard deviation rather than mean standard deviation for a global estimate of the experimental error.

3. Determine the number of degrees of freedom based on the total number of measurements in the data, i.e., equation (18).

References

1. Barlow RJ. 1989. Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences. 1st ed., Chicester: John Wiley and Sons.
2. Freund RJ, Wilson WJ. 2003. Statistical Methods. 2nd ed., Amsterdam: Academic Press.
3. Taylor JR. 1997. An introduction to error analysis : the study of uncertainties in physical measurements. 2nd ed., Sausalito: University Science Books.