H63 The Importance of Testing and Understanding Statistical Methods in the Age of *Daubert*: Can FORDISC Really Classify Individuals Correctly Only One Percent of the Time?

Nicole D. Siegel, DVM, Cleveland Museum of Natural History, 1 Wade Oval Drive, Cleveland, OH 44106-1767; and Stephen D. Ousley, PhD, Mercyhurst College, Department of Applied Forensic, Anthropology, 501 East 38th Street, Erie, PA 16546*

After attending this presentation, attendees will avoid future misunderstandings in the use of FORDISC and will be better able to use the program correctly and effectively.

This presentation will impact the forensic science community by exploring common misunderstandings in statistical analysis, particularly FORDISC.

Fordisc 3.1 (Jantz and Ousley 2005) uses discriminant function analysis, as has previous versions, and FORDISC has provided more and more additional information in addition to the classification results during its evolution. Failure to understand this additional information has led to a claim that challenges the accuracy of FORDISC. The recent publication of a series of FORDISC tests by Elliott and Collard (2009) is a result of their failure to appropriately interpret statistical results.

Elliott and Collard classified individuals from five groups in the Howells database (Berg, Northern Japan, Santa Cruz, Tasmania, and Zulu) into all Howells groups. They used all 56 craniometric variables in FORDISC as well as three groups of 10 variables from different cranial regions (basicranium, neurocranium, and face). Due to a misunderstanding of posterior probabilities, they reported very low percentage correct classification in general and concluded that FORDISC classifies less than 1% of individuals correctly. Their criterion was that classifications showing a typicality probability of less than 0.01 or a posterior probability of less than 0.8, no matter which group was most similar, were considered incorrect. Unlike typicality probabilities, the posterior probability does not have a threshold requirement. Higher posterior probabilities generally reflect higher probability of correct classification, but there are no specific recommendations or discrete cut-off values. In the statistical literature, having a posterior probability of at least 0.8 is merely considered a "strong" classification. Their test conditions seem rigged for failure: when using 56 variables, they were using more variables than many of Howells sample sizes, resulting in lower typicality probabilities, and when using only 10 variables form certain areas of the cranium, they were extremely unlikely to get high posterior probabilities. Additionally, they classified Howells individuals from the five groups using every other Howells group to ascertain if groups showed geographic similarity. However, they designated only one specific group from each region that should theoretically be the most similar one, and any other classifications were deemed incorrect. For instance, in Europe, only a classification of Howells' Berg individuals into Norse was considered correct.

Elliott and Collard's methods were followed as closely as possible using both FORDISC 3.1 (2005) and SAS 9.1 (2003), using the Howells database. Individuals from the same five ancestral groups were used, and run against all individuals in the Howells database. The analyses were conducted with all 56 variables and the same three groups of 10 variables representing the basicranium, neurocranium, and face. Because Elliott and Collard did not state which typicality probability was used, the chi-square typicality was used in this study. Correct analysis of the results resulted in correct cross-validated classification percentages of 18 to 32%, which is significantly greater than random, and greater than 1%. Classifications with higher posterior probabilities showed higher correct percentages, and regionally patterned variation was strongly indicated. The disparity between Elliott and Collard's conclusions and those of the current study is clearly a result of their misuse of posterior and typicality probability thresholds. Further, the geographic affinities of the test groups were confirmed when a more flexible criterion of regional similarity was accepted. Unlike Elliott and Collard's results, the current study showed that when the reference group is excluded, the percentage correct regional classifications is comparable to or slightly higher than the percentage correct classifications when the group is included in the analysis.

With the advent of *Daubert* standards, it is critical for forensic anthropologists to validate methods. The current analysis has shown that it is imperative to thoroughly understand the statistical underpinnings of any method, and that faulty criteria and test procedures can lead to false conclusions of low validity for a method. The number of measurements and stipulations for classification correctness used by Elliott and Collard resulted from a statistical misunderstanding that virtually guaranteed a low classification accuracy rate.

FORDISC, Discriminant Function Analysis, Statistics

 ** Presenting Author*