Correspondence

Commentary on: McBride DG, Dietz MJ, Vennemeyer MT, Meadors SA, Benfer RA, Furbee NL. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. J Forensic Sci 2001;46:424–428.

Sir:

In a recent article in this journal, McBride et al. (1) present a ground breaking forensic anthropological analysis based on an early "machine learning" algorithm. While there is much to commend in this article, as a first application there are also the inevitable points that raise questions relevant to future work in forensic anthropology. In this letter we raise a few of these issues.

Typically there are two goals to forensic anthropology. The first goal is that of estimation, in which a profile is built from the skeletal remains of an unidentified individual in the hopes that said remains may eventually be identified. As a statistical pursuit, such estimation needs to be probabilistically based. It is not enough, and indeed it is misleading, to state that the remains belonged to a "white male who was 42 years old and 6 feet tall." Continuous variables (such as age and stature) need to be stated as "highest posterior densities," while categorical variables (such as "race" and sex) should be given with their posterior probabilities. As the American public continues to be fed a diet of both reality-based and fictionalized accounts of detective and forensic anthropological cases, it is important that we provide information couched in the ambiguities of the science. Failure to do so may lead to missed opportunities at identification.

The second goal of forensic anthropology is to provide statistical evidence in the case of a putative (or so-called "positive") identification. This again is a probabilistic problem, as the anthropologist (minimally) needs to present a likelihood ratio (2). While there are many expert forensic anthropologists who are quite good at providing unbiased and efficient point estimates and categorical statements, if more "objective" forms of data (such as DNA) go challenged in the courts the day cannot be far off when an expert forensic anthropologist will face similar challenges on presenting evidence in an identification-based case. Forensic anthropologists who have had considerable training and much prior experience can, and often do, learn to glean a surprising amount of information from a single skeleton and its context. Machines (i.e., computers) are not particularly good in such situations, and it is doubtful that a jury will follow an argument based on "artifical intelligence" or presented on behalf of a "learned machine." Neither do we expect that judges will relish instructing a jury on how to interpret evidence presented on behalf of an "expert system," rather than by an expert.

All of this is not to say that computers have no place in the future of forensic anthropology. It is precisely in the domain of probabilistic statements that computers can be "taught to think" (i.e., programmed) quite effectively, while humans are understandably poor at processing the often massive amounts of data needed to calculate probabilities. As an early computer algorithm designed to solve discrete problems, the ID3 algorithm that McBride et al. used is not well suited to providing probabilistic statements. Newer statistical methods that bear some similarities to ID3, such as Bayesian CART (for "classification and regression trees") (3) can provide probabilistic statements. There are also now a large number of alternatives to ID3 (4), and older parametric models such as probit are also a possibility (5). And while we would not relish explaining Bayesian CART or cumulative probit models to a court, we would feel better about doing so than we would providing a "cut-and-dry decision" from ID3. A further problem with ID3, which is alluded to by McBride et al. ("ID3's prioritization of attributes may be viewed as problematic," p. 428) is that it does not tend to "think" in a manner similar to how expert forensic anthropologists analyze cases. ID3 is a so-called "greedy" algorithm, in that it partitions data sets by sequentially moving through attributes in order of information gain. In contrast, forensic anthropologists tend to use a "weight of information" approach, where they form opinions about cases based on agreement between a number of different attributes. In the following, we describe a recent large "experiment" in estimation of sex from the os coxae as a contrast to McBride et al.'s study.

In June of 1998 we scored 793 os coxae from the Terry Collection (the source for McBride et al.'s 115 cases) as part of a larger study on estimation of age-at-death from skeletal remains (NSF SBR-9727386, see http://konig.la.utk.edu/paleod.html). We describe our examination and scoring methods in detail here, as they are pertinent to how we made determinations of the sex of individuals. Because we collected pubic symphyseal ageing data that is scored differently for the sexes (6,7), it was necessary to "know" the sex of each individual when we were scoring. However, in an actual forensic context sex typically would not be known, and so we estimated the sex for each of the 793 cases using the three Phenice (8) characteristics that also appear in McBride et al.'s analysis. We scored each characteristic as "F," "F?," "?," "M?," "M," or unobservable. Now there is a bit of fiction involved in saying that we used the Phenice characteristics (and *only* those characteristics) in making determinations of sex. All os coxae were scored blind and independent of any other bones from the skeleton, but like in McBride et al.'s study the observer had access to the entire os coxae, and so may have used other (unrecorded) criteria. That said, we should also point out that our study was done principally over eight days, with the least number of os coxae scored on any one of these days being 49, and the most being 152. In addition to scoring the Phenice characteristics we were collecting pubic symphyseal and auricular phase data, as well as age "indicator" data from the cranium and long bones. We consequently spent considerably less time with each skeleton than would be true in a forensic context, and so it is unlikely that we used anything more than the Phenice characteristics with any regularity. In order to expedite the study we divided tasks so that the second author, who had the most experience in scoring auricular surfaces, scored most of the os coxae. Consequently, he was responsible for sexing about 90% of the bones, while the other authors each did about 5% (for those wanting more detail, the complete data set is available by anonymous ftp from the web-site listed above).

Instead of using ID3 with the Phenice characteristics we used a more modern program, Polytomous Logistic regression trees with Unbiased Split (PLUS) (9) to form a decision tree from the three

Phenice characteristics. PLUS is currently available from http://recursive-partitioning.com/plus/. We treated the characters as numeric, with "F" = 1, "F?" = 2, "?" = 3, "M?" = 4, and "M" = 5. Unobservable traits were handled using "nodewise imputation" and for cross-validation we used individual cases (i.e., a "leave-one-out" strategy). The program found that the lowest cross-validated error rate occurred when there was a simple split on subpubic concavity, with "F" and "F?" going to the "left" and "?," "M?," and "M" going to the "right." This simple tree gave a cross-validated error rate of 2.4%, misclassifying 8 of the 361 actual females in the sample as males and 11 of the 432 actual males as females. This misclassification rate is lower than all of the "mean error percents" given in McBride et al., most of which were found using a large number of attributes, as versus the single attribute subpubic concavity. While again, it could be argued that our scoring of subpubic concavity was influenced by observing other unrecorded attributes in the os coxae (and was not made independent of the other two Phenice characteristics), such an argument cannot be made for a study by Sutherland and Suchey (10). In a study of 1284 pubic bones that had been removed at autopsy, and in which only Phenice's "ventral arc" could be scored for its presence/absence, the authors misclassification rate was 4%, again lower than all of McBride et al.'s mean rates.

Because decision trees from PLUS, like those from ID3, are univariate, we also consider the "Linear Machine Decision Tree" (LMDT) algorithm. LMDT (11) uses a multivariate model (discriminant analysis) in a decision tree setting, and allows for "pruning" of the tree so that it does not become overly complex ("bushy"). The source code for LMDT is currently available from http://yake.ecn.purdue.edu/~brodley/software/lmdt.html. We applied LMDT, again using node-wise imputation for missing data. Ultimately, LMDT misclassified 9 of the 361 actual females as male (one more misclassification than PLUS) and 6 of the 432 actual males as female (5 less than PLUS with cross-validation, or 4 less than PLUS using the biased "plug-in" rule). For complete cases (i.e., cases with no missing data) LMDT first makes a split simultaneously on the basis of the ventral arc and sub-pubic concavity. Cases where the sub-pubic concavity is scored as "F" and the ventral arc as no more male than "M?," or the sub-pubic concavity is scored as "F?" and the ventral arc as no more male than "?," or the sub-pubic concavity is scored as "?" and the ventral arc as no more male than "F" are all classified as "female." The remaining cases are then split again using the sub-pubic concavity and ischio-pubic ridge. Cases that have both features scored as "M" are classified as male. Cases that do not have both features scored as "M" go through further splits, but in the interest of brevity we do not describe them here.

We can also compare our results using decision trees to what we obtained simply by "mentally processing" the information from the Phenice characteristics. Some points are in order here before we turn to this comparison. First, there were three clerical errors we found after the fact. We correctly classified two individuals by sex, even though we have no recorded Phenice characteristics from these individuals. As we have scores for their pubic symphyseal development it is likely that we simply failed to record the Phenice characteristics for these two individuals. There is an additional clerical error, in that one individual for whom we scored all three characteristics as unambiguously female we also have classified as a male. Outside of these three errors, there are four cases where our opinion on sex classification is not consonant with the Phenice characteristics, although in all four instances we correctly identified the sex. In one actual male, where we scored the ventral arc as "F," the subpubic concavity as "F?," and the ischio-pubic ridge as "M" we ultimately classified the individual as male. For two actual males where we scored the ventral arc as "M" and the other two characters as "F?" we identified the individuals as males. Finally, for one actual male where we scored the ventral arc as "F?," the subpubic concavity as "M?," and the ischio-pubic ridge as "?," we ultimately identified the individual as a male. In all four cases, we must presume either that we were drawing on other attributes, or that our visual weighting is not well reflected in the three Phenice scores. Aside from these exceptions, all of the sexes we assigned in June, 1998 were done by taking the majority characteristic. In other words, if two of the characteristics were "F" while one was "M" or "?" we classified the individual as a female, and we treated unobservable characteristics as uninformative. On this basis, we ultimately misclassified 7 of the 361 actual females in the sample as males (2 less than LMDT) and 5 of the 432 actual males as females (1 less than LMDT).

Now, how might we use the above-described study in forensic settings? Our "mental processing" did not use the sample itself to derive rules, so our observed misclassification rates should be applicable to any new samples. By Bayes Theorem, if we assume that an unidentified case is as likely to be from a male as from a female, then upon "determining" sex on the basis of the Phenice characteristics we should say that individuals we identify as female have a 0.9983 posterior probability of actually being female (equal to (354/361)/(354/361 + 5/432)). Similarly, those we call "male" have a 0.9808 posterior probability of being male. In an identification case, if the identification (external to the osteological evidence) is for a female, and we suggest on the basis of the Phenice characteristics that the individual has the morphology of a female, then the likelihood ratio (assuming an even sex ratio in the "population at large") is 1.9667, while for a male who has male Phenice characteristics the likelihood ratio is 1.9615. While these are very weak likelihood ratios (because there are only two sexes, which we have assumed are equally frequent), they could be combined with likelihood ratios from other osteological evidence in order to "sharpen" the posterior odds. That these likelihood ratios are nearly 2.0 is because the level of misclassification is extremely low. "Machine learning" methods could presumably be applied to other osteological classification problems, where the number of categories is greater than two. While such classifications would be more informative in identification cases, we also suspect that the level of misclassification could be unacceptably large. But for the present setting of sex identification from the os coxae, we do not find that "machine learning" methods offer much beyond what a reasonably well-trained and experienced osteologist can provide. In point of fact, McBride et al.'s ID3 based analysis did not perform nearly as well as we did, calling into question not only the algorithm, but also the quality of the attribute scores from which the program generated its decision tree.

References

1. McBride DG, Dietz MJ, Vennemeyer MT, Meadors SA, Benfer RA, Furbee NL. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. J Forensic Sci 2001;46:427–31.
2. Evett IW, Weir BS. Interpreting DNA evidence: statistical genetics for forensic scientists. Sunderland, MA: Sinauer Associates, 1998.
3. Chipman H, George EI, McCulloch RE. Bayesian CART model search. J Am Statist Assoc 1998;93:935–60.
4. Lim T-S, Loh W-Y, Shih Y-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine learning 2000;40:203–28.
5. Konigsberg LW, Hens SM. Use of ordinal categorical variables in skeletal assesment of sex from the cranium. Am J Phys Anthropol 1998;107: 97–112.
6. Gilbert BM, McKern TW. A method for aging the female os pubis. Am J Phys Anthropol 1973;38:31–8.

7. McKern TW, Stewart TD. Skeletal age changes in young American males. Quartermaster Research and Development Command Technical Report EP-45, 1957.
8. Phenice TW. A newly developed visual method of sexing the os pubis. Am J Phys Anthropol 1969;30:297–302.
9. Lim T-S. Polytomous logistic regression trees (dissertation). Madison (WI): Univ. of Wisconsin, 2000.
10. Sutherland LD, Suchey JM. Use of the ventral arc in pubic sex determination. J Forensic Sci 1991;36:501–11.
11. Brodley CE, Utgoff PE. Multivariate decision trees. Machine learning 1995;19:45–77.

Lyle W. Konigsberg, Ph.D.
Department of Anthropology
University of Tennessee
Knoxville, TN 37996-0720

Nicholas P. Herrmann, M.A.
Department of Anthropology
University of Tennessee
Knoxville, TN 37996-0720

Daniel J. Wescott, M.A.
Department of Anthropology
University of Tennessee
Knoxville, TN 37996-0720

Authors' Response

Sir:

We are pleased to have generated interest in our recent paper (1). Konigsberg, et al. (2) present a valuable contribution to the use of machine learning algorithms in questions relevant to forensic science. Their work points out the increasing importance of presenting the conclusions of osteological analyses in probabilistic terms, but the direct criticisms of our work distort its central concept.

Our study (1) proceeded from the point of view that, in order to obtain an accurate and informative analysis of skeletal remains, one must first decide what, specifically, to analyze. Expert skeletal analysts have the benefit of extensive training and experience to guide and support their conclusions, and will have refined their own preferred techniques over many years. In contrast, skilled, but nonexpert, analysts often face difficulties when compelled to select among many methods and criteria, as may be the case with, for example, fragmentary or otherwise unusual material. Phenice (3), Sutherland and Suchey (4), Rogers and Saunders (5), and most recently, Konigsberg, et al. (2), show that small groups of attributes can be quite accurate. Small sets of attributes have an added advantage over large sets in being less likely to include unobservable attributes. However, no attribute set is necessarily optimal in every instance.

McBride, et al. (1) presented ID3 in a useful, nonarbitrary and repeatable procedure for identifying good subsets of attributes. To show that the chosen attributes worked as well as all attributes together, training sets of 70% of the sample were drawn randomly with replacement and tested against the remaining 30%. Relative error rates averaged over ten trials were cited for each attribute suite as proof of success in this regard. We stated that "The selected attribute suite of preauricular sulcus, sciatic notch, and subpubic concavity should provide good results when scored as indicated in Table 1," but did not state a specific level of accuracy that one should expect in a different context. Because we developed the attribute suites with a bootstrapped training set/test set protocol and presented ten-trial averages for each attribute set, we regard our results as robust.

The 31 attributes used in the study were based on the techniques of three widely recognized experts in skeletal analysis (see Table 1 (1)). They are therefore partly redundant in that some attributes require different criteria for scoring the same os coxae features. ID3 demonstrated sensitivity to semantic differences in the definitions of each attribute. Out of approximately 90 trials, some "versions" of attributes were never preferentially chosen by ID3, and others commonly were. Similarly, several attribute descriptions commonly appeared as second or third branches on decision trees; never first, never last, suggesting those attributes contained information useful in segregating cases that remained ambiguous after the first split. The sensitivity of ID3 to these subtleties may reflect intraobserver variability in applying each criterion, or it may suggest a broader interpretation, for example, that Expert C's "subpubic concavity" criteria capture more information than Expert B's. If taken in context, it does not necessarily, as Konigsberg, et al. state, "call into question not only the algorithm, but also the quality of the attribute scores from which the program generated its decision tree." (2). Instead, it suggests that ID3 may be useful in examining relative effectiveness of individual attributes, questions of inter- and intraobserver variability, and the semantics of attribute definitions.

To address the question of accuracy with results more directly comparable to Konigsberg, et al. (2) than our original study presented, we obtained the larger Terry Collection data set made available on the Internet by Konigsberg, as well as the LMDT (6) and PLUS (7) software (see Konigsberg, et al. (2) for URLs). We also ran new trials using ID3 with our original, raw sample of 115 Terry Collection individuals. Using a "leave-one-out" jackknife procedure with our sample of 115 (35F, 80M), we tested our preferred attribute suite of preauricular sulcus, sciatic notch, and subpubic concavity (1), as well as attribute sets we designated to represent those recommended by Phenice (1,3). We also intended to test our 115 sample and one training set/test set trial under PLUS and LMDT. However, LMDT has not run successfully on our system, and technical support is no longer offered by its author (Carla Brodley, e-mail communication, 2001); results are given only for PLUS. Finally, using ID3 we tested 600 individuals (total 300F, 300M) from the Konigsberg Terry Collection sample. It was necessary to divide the sample of 600 into three trials of 200 each (100F, 100M) because our implementation of ID3 is an MS-DOS program that can only access enough memory for slightly more than 200 cases at once. We must stress that this is a limitation of our ID3 software, not of the algorithm itself.

Using the Terry Collection data provided by Konigsberg (Phenice's attribute set) and a leave-one-out protocol, ID3 misclassified a total of 8/300 females (2.6%) and 7/300 males (2.3%), for an overall cross-validated error rate of 15/600, or 2.5%. In one trial using our 115 sample with ID3, our preferred suite of attributes misclassified 3/35 females and 1/80 males, for a total error rate of 4/115 (3.4%). The "Phenice" attributes designated in our original paper (a "generic" set, comprised of one attribute definition from each expert; see Table 1 (1)) misclassified 8/35 females and 4/80 males. For comparison with our generic "Phenice" suite, we re-ran the trial three times, each time using the three "Phenice" attributes as defined by each expert. Interestingly, misclassification rates were slightly different for the various definitions of the "Phenice" attributes: Expert A's suite misclassified 8/35 females and 2/80 males, Expert B's suite misclassified 8/35 females and 4/80 males, and Expert C's suite misclassified 5/35 females and 1/80 males. Each trial with the "Phenice" attributes generated a distinct, albeit similar, tree in which cases were ordered differently and some different cases were misclassified, despite having been developed from the same data. This further suggests that ID3 is sensitive to subtle, qualitative differences in scoring, as noted above.

We tested PLUS using our 115 sample and one arbitrarily chosen training set/test set sample that was originally drawn for the

ID3 trials. With 115 cases, our preferred suite of attributes misclassified 1/35 (2.8%) females and 4/80 (5%) males. The "Phenice" attributes misclassified as follows: the generic attributes misclassified 3F and 9M; the Expert A attributes misclassified 10F and 2M; the Expert B attributes misclassified 12F and 3M; the Expert C attributes misclassified 5F and 2M. In the training set/test set trial with PLUS, our preferred set of attributes misclassified one female and two males, with a cross-validated error rate of approximately 2.5%. The generic "Phenice" attributes misclassified eight females and one male, with a cross-validated error rate of 11.2%.

The brief comparisons presented above do not suggest that ID3 is inferior to PLUS. Although our data sets returned slightly weaker results in some trials, it should be noted that the ossa coxarum were originally scored by one individual with the specific objective of duplicating the recommendations of three different experts in skeletal analysis. That ID3 and PLUS both detected subtle differences among attribute suites (reflected in varying accuracy rates for the "Phenice" trials) suggests the scorer's original objective was met. In regard to our "poor" results with ID3 and the Phenice characteristics, we are not surprised that Konigsberg, et al. (2) obtained a better result using a sample nearly seven times larger. Methods do better on the samples from which they are developed. Phenice (3) obtained his results using 275 individuals from the Terry Collection. Our results (1) were obtained from a Terry Collection sample of 115 individuals and validated with a rigorous training set/test set protocol averaged over 10 trials. Since it is quite likely that substantial portions of our and Phenice's samples are contained within the Konigsberg Terry Collection sample, we would be concerned if Konigsberg, et al. (2) had not obtained a better result. Under the more directly comparable protocols presented here, the misclassification rates of 2.6% F and 2.3% M for ID3 using the Konigsberg Terry Collection sample evaluate favorably with the errors of 2.2% F and 2.5% M (PLUS), and 2.4% F, 1.3% M (LMDT) reported by Konigsberg, et al. (2).

We agree with Konigsberg, et al. (2), that there is a need to develop probabilistic statements for categorical variable estimations, such as sex. Either ID3 or PLUS should be useful in these and other applications of machine learning algorithms to skeletal analyses. Several other algorithms and/or software based upon them were noted by Konigsberg et al. (2) and McBride, et al. (1). They are also readily available on the Internet and merit evaluation in future studies. It is important to note in closing that none of these programs constitute expert systems in themselves. They are tools intended to assist the development of rule sets, or decision trees, such as those discussed above, which are the building materials of expert systems. A finished expert system may incorporate several hundred rule sets, much refined after consultation and validation with several human experts. Such a hypothetical system would produce skeletal analyses equivalent to analyses done by the experts who contributed to its development.

References

1. McBride DG, Dietz MJ, Vennemeyer MT, Meadors SA, Benfer RA, Furbee NL. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. J Forensic Sci 2001;46:427–31.
2. Konigsberg LW, Herrmann NP, Wescott DJ. Commentary on McBride DG, Dietz MJ, Vennemeyer MT, Meadors SA, Benfer RA, Furbee NL. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. J Forensic Sci 2001;47(2):000–000.
3. Phenice TW. A newly developed visual method of sexing the os pubis. Am J Phys Anthropol 1969;30:297–302.
4. Sutherland LD, Suchey JM. Use of the ventral arc in pubic sex determination. J Forensic Sci 1991;36:501–11.
5. Rogers TL, Saunders SR. Accuracy of sex determination using morphological traits of the human pelvis. J Forensic Sci 1994;39:1047–56.
6. Brodley CE, Utgoff PE. Multivariate versus univariate decision trees. Amherst (MA): University of Massachusetts, Department of Computer and Information Science; COINS Technical Report No. 92-8, 1992.
7. Lim T-S. Polytomous logistic regression trees [dissertation]. University of Wisconsin, 2000.

David G. McBride, M.A.
Robert A. Benfer, Jr., Ph.D.
Department of Anthropology
University of Missouri
Columbia, MO 65211

Commentary on: Dou C, Bournique J, Zinda M, Gnezda M, Nally A, Salamone S. Comparison of Rates of Hydrolysis of Lorazepam-Glucuronide, Oxazepam-Glucuronide and Temazepam-Glucuronide Catalyzed by E. Coli β-Glucuronidase Using the Online Benzodiazepine Screening Immunoassay on the Roche/Hitachi 917 Analyzer. J Forensic Sci 2001;46(2):335–340.

Sir:

In their article, Dou, et al. report catalytic rates and Km values for the hydrolysis of lorazepam-glucuronide, oxazepam-glucuronide, and temazepam-glucuronide by *E. Coli* β-glucuronidase. The authors report that the purity of the glucuronide reference materials, purchased from Alltech, is greater than 90%; however, they do not mention if they actually validated the concentrations of these reference materials prior to use or merely reported the manufacturer's labeled concentration. Although Alltech lists the concentration of these reference materials as 1 mg/mL in their catalog, they do note that due to limited supplies of pure standard, these solutions are not quantitative.

In a previous study (2) we reported on the validation of (R,S) lorazepam- and oxazepam β-glucuronide primary reference materials for hydrolysis and quality assurance controls. Using HPLC analysis and GC/MS analysis following both acid and β-glucuronidase hydrolysis, we found that the lorazepam β-glucuronide material was >95% pure, but the oxazepam β-glucuronide material was only 54% pure. At the time the manufacturer stated that their reference materials were highly purified (>95% pure) but were intended for use only as qualitative standards. The materials were semi-quantitative and were not intended for use as quantitative standards. In order to perform the cross reactivity and kinetic studies described by Dou, et al., the concentration of the substrates must be known in order to obtain valid results. We are concerned that the authors made no mention of the validation of the concentrations of these glucuronide reference materials prior to performing these studies. Without the validation information for the reference material purity data, it is unclear whether the differences in the hydrolysis rates and Km values are a result of structural differences or an error in the substrate concentration.

Although the benzodiazepines are stereospecifically conjugated so that only one isomer (the β-glucuronide) exists in biological systems, both oxazepam and lorazepam are racemates since the 3 carbon on the diazepine ring is chiral. Therefore, both the R and S isomers of the β-glucuronide are present in urine, plasma as well as these reference materials. Depending on the source of the β-glucuronidase, the S and R isomers can be hydrolyzed at different rates (3). It has been reported that the β-glucuronidase from *E. Coli* is very selective hydrolyzing the S isomer 446 times faster than the R isomer (at 37°C). Therefore, the relative concentrations of the S and R isomers may impact the initial hydrolysis rates of these reference materials. We reported that relative concentrations of the S and R isomers of oxazepam β-glucuronide were 61.0% and 39.0%, respectively and 54.5% and 45.5% for lorazepam β-glucuronide (2). It is not known if the relative concentrations of the isomers in the reference materials have changed over time.

As stated in our previous study, Alltech's glucuronide materials can be valuable as hydrolysis controls in method development and routine benzodiazepine analyses, but each laboratory must validate the purity of these glucuronide reference materials prior to use.

References

1. Dou C, Bournique J, Zinda M, Gnezda M, Nally A, Salamone S. Comparison of Rates of Hydrolysis of Lorazepam-Glucuronide, Oxazepam-Glucuronide and Temazepam-Glucuronide Catalyzed by E. Coli β-Glucuronidase Using the On-line Benzodiazepine Screening Immunoassay on the Roche/Hitachi 917 Analyzer. J Forensic Sci 2001;46(2):335–40.
2. O'Neal C, Poklis A. Validation of benzodiazepine β-glucuronide primary reference materials for hydrolysis and quality assurance controls. Forensic Sci Int 1996;79:69–81.
3. Reulius H, Tio C, Knowles J, McHugh S, Schillings R, Sisenwine S. Diastereoisomeric glucuronides of oxazepam. Drug Metab Dispos 1979;7(1):40–3.

Additional information and reprint requests:
Carol L. O'Neal, Ph.D.
Toxicology Laboratory
Division of Forensic Sciences
Fairfax, VA

Alphonse Poklis, Ph.D.
Medical College of Virginia Campus
At Virginia Commonwealth University
Richmond, VA 23298-0165

Authors' Response

Sir:

The authors O'Neal and Poklis bring up valuable points that are thoroughly discussed in their paper (*Forensic Science International* 79, (1996) 69–81). They question whether or not validation of the glucuronide reference materials was performed prior to the experiments conducted to determine the catalytic rate constants for the hydrolysis of lorazepam-glucuronide, temazepam-glucuronide, and oxazepam-glucuronide. We wish to acknowledge the fact that although no in-house analytical determination for purity of the glucuronides was performed, we did obtain the analytical Standard Certificates of Quality from Synthetic Technology Corp (manufacturer of the material). The report by O'Neal and Poklis in 1996 induced Synthetic Technology Corp to reexamine their methods of preparation and characterization of the benzodiazepine glucuronide material.

The reported purity was >90%, >95% and >95% for temazepam-glucuronide, lorazepam-glucuronide, and oxazepam-glucuronide, respectively. This was based on HPLC analysis. In addition (precipitated by the concerns of O'Neal and Poklis) we contacted the vendor and obtained the following information[1]:

- The temazepam-glucuronide used in the studies was initially bottled in August 1993 from stock synthesized in April of that year. It was the vendor's first synthetic batch of this material. The vendor used 7.0 mg of exhaustively purified (by NMR) stock and diluted by weight into HPLC grade methanol to a resulting concentration of 0.1mg/mL. The chromatogram supplied with the material was checked and found to be accurate, indicating the primary "contaminants" being the minor "satellite" sugar enantiomers of the R/S mix. The stated purity for this lot of material was given as greater than 90%.
- The Lorazepam Glucuronide used in the studies was prepared in November 1995 as a part of a lot of 108 ampoules. It was prepared in a similar manner as described above with 10.9 mgs of exhaustively purified stock diluted in HPLC grade methanol and immediately sealed in silylated vials. Both the purity and the calculations of concentration have been verified to be correct.
- The oxazepam glucuronide used in the kinetic studies was prepared in December 1997 as a part of 108 ampoules. 11.74 mgs of exhaustively purified oxazepam glucuronide stock was diluted in HPLC grade methanol and immediately sealed in silylated vials. Again, both the purity and the calculations of concentration were reviewed and determined to be accurate. This material is probably the purest of all three.

Of the three lots the Temazepam is the oldest material, of the lowest (>90%) original purity at the time of shipment. The "primary" impurity was determined to be the satellite mix of enantiomers based upon the four possible combinations of the coupling of the sugar with the hydroxylated benzodiazepine.

The biggest concern we had regarding the purity of the material was with the stability of the glucuronidated conjugate. Upon aging and degradation the free aglycone, or similar chemical byproducts would be expected to be present. We tested this purity by antibody cross-reactivity. If there was a substantial amount of non-glucuronidated material, the immunoassay (which has low selectivity to the glucuronides and high cross-reactivity to the free drug) would have picked it up. Our study indicated that the lots did not contain significant amounts of this expected contaminant.

The main focus of the study was to compare hydrolytic rates of several benzodiazepine-glucuronide conjugates under various conditions. If the material was not pure the catalytic turnover (Kcat) and effects of temperature, matrix and pH on Kcat would not significantly change the values since the substrate is used in excess and only product formation is monitored. The Km would, however, be effected and conclusions concerning the binding affinity of each glucuronide to the enzyme would be in question.

O'Neal and Poklis also commented on the fact that depending on the source of the β-glucuronidase, the rates of hydrolysis of the R and S isomers vary. They state that β-glucuronidase from *E. coli* hydrolyzes the S isomer 446 times faster than the R isomer (at 37ºC). In the case of our studies we were measuring initial rates (two minute reaction) so that the selectivity for one isomer over the other would not be observed. The enzyme may have been more selective for the S isomer and if so the resulting product would have been the corresponding enantiomer. In the light of this selectivity it would have been better to use the pure diastereoisomers for this study.

The authors raise good points about the chemical and stereochemical purity of the glucuronide conjugates. While we feel confident of the purity provided by the information from the manufacturer and by the testing that we did in house, it still would have been prudent (in light of the O'Neal and Poklis paper) to additionally test for the purity using another physical method.

Reference

1. Benzodiazepine glucuronide standards information was obtained from Dr. Mark Hagadone (Synthetic Technology Corp.) via personal communication.

Matt Gnezda
Roche Diagnostic Corporation

Salvatore J. Salamone
Advance BioTech Consulting