A Comparison of Two Versions of the Biofidelity Ranking System (BioRank): Implications on THOR-50M and THOR-05F Biofidelity Evaluations

Devon L. Albert

Abstract The Biofidelity Ranking System has been implemented in a variety of studies since its inception in 2002 to assess the biofidelity of anthropomorphic test device and model responses against a human response. Four chronological versions have been developed with each new version altering how biofidelity scores are calculated. However, it is unknown if and how differences in Biofidelity Ranking System versions affect the resulting biofidelity scores and the conclusions of the studies that implement them. Therefore, the objective of this study was to compare the biofidelity scores of the 2009 and 2018 versions of the Biofidelity Ranking System by applying both versions to the same dataset. The dataset in this study consisted of matched Hybrid III and THOR 50th percentile male sled tests and included 101 response signals from each anthropomorphic test device. The results of the analysis indicated that the 2018 version produced significantly lower (better) scores for individual response measurements than the 2009 version and a significantly better overall biofidelity assessment, depending on how the individual response scores were averaged. This finding implies that using different versions of the Biofidelity Ranking System on the same dataset may affect the biofidelity results and conclusions. Based on the results of this study, recommendations for transparently using the Biofidelity Ranking System are provided.

Keywords Anthropomorphic test device, biofidelity, human response corridors, objective rating metric, THOR.

I. INTRODUCTION

The Biofidelity Ranking System (BioRank) was first developed in 2002 to assess the biofidelity of the WorldSID-α prototype anthropomorphic test device (ATD) relative to other side impact ATDs [1]. BioRank has since been used to assess the biofidelity of various ATDs [1-6], ATD components [7-14], and finite element and multi-body models [15-23]. One advantage BioRank has over other objective rating metrics, such as CORA [24] or ISO/TS 18571 [25], is that it was specifically developed to assess biofidelity as opposed to similarity. Since CORA and ISO/TS 18571 assess similarity between multiple responses, biofidelity can only be assessed via comparative biofidelity. Specifically, two ATDs or models must be evaluated against the same response targets so that the resulting objective rating scores for each ATD or model can be compared. Then, the ATD or model with the best score can be considered the most biofidelic. However, BioRank can compute a standalone biofidelity score for a single ATD or computational model.

Four versions of BioRank have been developed with differences in the implementation of each variant [1][26-28]. All versions of BioRank were proposed in the context of comparing ATD response time histories to target human response time history corridors. In the context of this paper, *human* is defined as human volunteers or post-mortem human subjects (PMHS). The first version of BioRank (BioRank 2002) also introduced a method of evaluating peak responses of ATDs in the absence of human time history corridors [1]; however, that methodology has not been evaluated in this study. A summary of each version of BioRank follows, but a detailed explanation of each version can be found in the Appendix, and a tabular summary of each version is provided in Table AI.

BioRank 2002 introduced a two part system that first assessed the biofidelity of individual ATD response measurements, and then combined those measurements via an averaging scheme to assess overall ATD biofidelity. To generate BioRank scores for each response measurement, the variance between the ATD and mean human response curves were normalized by the variance within the human response. Therefore, a lower score was more desirable, and the magnitude of the score generally represented how many standard deviations the

D. L. Albert (phone: +1-540-231-1890, e-mail: dla16@vt.edu) is a Postdoctoral Associate in the Center for Injury Biomechanics (CIB) at Virginia Tech in Blacksburg, VA, USA.

ATD response was from the human response. After calculating the biofidelity score for each response measurement, a hierarchical averaging and weighting scheme was implemented to combine the scores of all response measurements into a single overall biofidelity score for the ATD. Details on the averaging scheme can be found in the Appendix, but it is important to note that it involved classifying response signals as external or internal and averaging each type of signal separately. The second version of BioRank (BioRank 2009) modified the definitions of internal and external signals used in the averaging scheme and removed the weighting system from the hierarchical averaging scheme [28]. Detailed definitions of internal and external responses within the framework of BioRank 2009 can be found in [29]. The third version of BioRank (BioRank 2013) made recommendations for data processing prior to calculating BioRank and modified the calculation of the response scores [27]. Specifically, phase optimisation was used to shift the ATD curve to better align with the human curve before the response scores were calculated. The individual response score was then modified to be combination of a phase score (P) and a shape and magnitude score (SM), where the SM score was calculated using the same equations as the individual response scores in previous versions of BioRank. Finally, the fourth version of BioRank (BioRank 2018) modified the phase optimization procedure and changed the equation used to calculate the SM score from a ratio of quadratic functions to a ratio of linear functions. [26]. It was also suggested in BioRank 2018 that it is not necessary to evaluate internal and external biofidelity separately, as in previous versions of BioRank, and that the two signal types could be grouped together within the hierarchical averaging scheme.

The way in which BioRank has been implemented in the literature varies widely. Several studies do not implement the hierarchical averaging portion of BioRank, presenting only the individual response scores and sometimes performing a non-hierarchical, unweighted average of all signals [7-16][18-22]. Often, this is because BioRank is being applied to a specific body region or component only; therefore, implementing the hierarchical averaging is not possible or feasible. Of the studies that do implement the hierarchical averaging, the most common point of deviation is in the division of responses into internal and external measures. Two studies conducted before BioRank 2018 did not differentiate between internal and external because the available responses were heavily skewed toward one type [4][6]. Additionally, [3] used BioRank 2009 to perform the biofidelity analysis in their study, but used some definitions of internal and external that were more consistent with BioRank 2002. Specifically, [3] classified body region displacements as external signals, while BioRank 2009 classified such signals as internal. Another potential area of deviation is in the development of the human response corridors. BioRank requires one standard deviation response corridors. However, how the standard deviation corridors are generated is open to interpretation. Standard deviation can be calculated for a population or a sample, leading to differing corridors widths that are more different at lower sample sizes. Additionally, standard deviation can be calculated in one or two dimensions, particularly for force-deflection curves where the standard deviation corridors are based on the variances in both the force and deflection [3][8][14]. Finally, the standard deviation of an existing corridor from the literature cannot always be calculated due to unavailability of the original individual responses [1][3]. It is currently unknown how the use of different human response corridors or different averaging techniques can influence the resulting BioRank score and the corresponding conclusions regarding biofidelity.

An interesting case study in the literature is the biofidelity evaluation of the THOR frontal ATDs. BioRank 2009 was used in 2017 to conduct the biofidelity analysis of the THOR 50th percentile male (THOR-50M or THOR) with respect to the Hybrid III 50th percentile male (HIII) [3]. The following year, the biofidelity evaluation of the THOR 5th percentile female (THOR-05F) was performed using BioRank 2018 [5]. As mentioned above and detailed in the Appendix, BioRank 2009 and 2018 use very different methods for calculating the biofidelity response scores. However, it is unknown how these differences affect the individual response scores and the overall biofidelity scores at the culmination of the hierarchical averaging scheme. Large differences in scores between different versions of BioRank may alter the conclusions of existing biofidelity studies in the literature, such as those for the THOR ATDs. Therefore, the purpose of this study was to compare the biofidelity scores of the 2009 and 2018 versions of BioRank by applying both versions to the same dataset. These two versions were chosen to for several reasons. First, BioRank 2009 and 2018 were used to evaluate the biofidelity of the THOR-50M and THOR-05F, respectively, making them of interest to compare. Second, BioRank 2009 is functionally equivalent to BioRank 2002 when the hierarchical averaging scheme is not implemented, and these two versions were the most used versions in the literature [2-4][7][8][10-12][14-23][30], especially among studies that did not implement the hierarchical averaging scheme [7][8][10-12][14-16][18-22]. Although BioRank 2013 has currently been

implemented more [6][9][13] than BioRank 2018 [5], BioRank 2018 is the most recent version and is expected to gain use as more time passes. Finally, BioRank 2018 clarifies some of the ambiguities associated with the implementation of BioRank 2013, such as the use of constant width corridors and the methodology for determining the optimal phase shift for the ATD response.

II. METHODS

The dataset used in this study to compare BioRank versions consisted of previously published matched frontal sled tests performed using the THOR-50M (THOR-NT with mod-kit, making it functionally equivalent to the THOR-metric), HIII 50th percentile male, and approximately 50th percentile male PMHS [31][32]. Both ATDs were certified prior to the test series. The tests were designed to replicate a Toyota Camry New Car Assessment Program (NCAP) test (ΔV = 56kph, peak acceleration = 470 m/s^2) and were conducted under three different restraint conditions: knee bolster (KB), KB and steering wheel airbag (KB/SWAB), and knee bolster airbag and SWAB (KBAB/SWAB). All conditions included a three-point seatbelt with a pretensioner and load limiter. Measured subject responses included linear accelerations in the X and Z directions, as defined by SAE J211 [33], and angular velocities in the X, Y, and Z directions for the head, thorax and pelvis. Excursions of the head, shoulders, hips, knees, and ankles in the X and Z directions were also quantified using motion capture. Finally, thoracic deflections were quantified at the upper sternum, upper left, upper right, lower left, and lower right locations on the thorax using chest bands. Further details on test methodology and individual response time histories can be found in [31][32].

Multiple tests were performed for each surrogate type within each test condition so average responses per condition were calculated before performing the BioRank analysis. Specifically, two tests were performed for each ATD and each condition. For the PMHS, two tests were performed for the KB condition and three tests for each of the KB/SWAB and KBAB/SWAB conditions. Characteristic averages were calculated for each surrogate within each test condition. For the PMHS tests, one standard deviation corridors were also calculated around the PMHS characteristic average using the population form of standard deviation. Before each response curve was used in BioRank 2009 or 2018, it was truncated to the relevant signal duration using the CORA truncation algorithm [24]. Variable settings for the truncation algorithm can be found in [31].

Some data in the original study were excluded from this analysis because the PMHS standard deviation corridors could not be calculated when only one subject response was available for a particular measurement within a particular test condition. These data included the pelvis angular velocity for the KB condition and the left and right hip excursions. After these exclusions were made, 101 response measurements were included in this study's BioRank analysis across all body regions and test conditions for each ATD.

BioRank Calculations

A custom Matlab (R2016b, MathWorks, Natick, MA, USA) script was written and implemented to calculate biofidelity scores according to BioRank 2009 and BioRank 2018. Both versions of BioRank were applied to the THOR and HIII responses in the dataset. This resulted in a total of 404 biofidelity response scores. The detailed steps to calculate the biofidelity response scores for each version are described in the Appendix.

The same hierarchical averaging scheme was used for both BioRank 2009 and 2018 in the current study. Although BioRank 2009 specifies separation of internal and external signals during averaging while BioRank 2018 does not, all of the signals in the current study would be considered internal according to both BioRank 2009 and BioRank 2018. As a result, there was no effective difference between the BioRank 2009 and 2018 averaging schemes for this dataset. A schematic of the hierarchical averaging scheme for this dataset is shown in Figure 1. First, all of the biofidelity scores for response measurements, e.g., acceleration in the x direction, acceleration in the z direction, etc., within a particular test condition and body region were averaged to obtain an average score for each test condition within a body region. Next, all of the average biofidelity scores for each test condition were averaged within a particular body region to obtain an average score for each body region. For this study, the test conditions were KB, KB/SWAB, and KBAB/SWAB. Finally, all of the average biofidelity scores for each body region were averaged to obtain an overall biofidelity score. For this study, the body regions were the head, thorax, pelvis, knee, and ankle. See Table AII in the Appendix for a list of all response signals and their associated body regions for this study's dataset. If both external and internal responses were evaluated, they would have been

divided on the response level such that two separate averaging *trees* would exist to create an overall internal biofidelity score and an overall external biofidelity score, according to the BioRank 2009 methodology. Then, an overall biofidelity score would have been calculated by averaging the internal and external overall scores. Equation A10 in the Appendix summarizes the averaging scheme used in the current study.

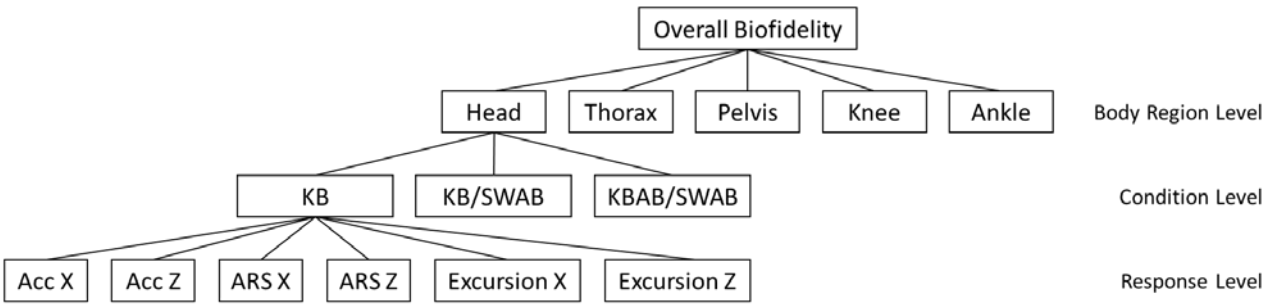

Fig. 1. Hierarchical averaging scheme used for the current study, where only internal responses were measured.

Data Analysis

The resulting biofidelity scores were compared between BioRank 2009 and 2018 using several methods. First, a Wilcoxon signed-rank test was used to compare the 101 pairs of BioRank 2009 and 2018 response scores within each ATD to evaluate whether the 2009 and 2018 scores were statistically different. The Wilcoxon signed-rank test was used instead of the paired t-test because the biofidelity response scores were very skewed and did not follow a normal distribution.

The differences between the overall BioRank scores generated from the hierarchical averaging scheme were evaluated using a critical difference analysis. The procedure for the critical difference analysis was developed in [28] to quantify a critical difference in overall BioRank values above which two ATDs could be considered significantly different. It was adapted for use in the current study to determine a critical difference between the overall BioRank 2009 and 2018 scores for each ATD. First, the response scores were averaged across conditions so that there was one average score for each response measurement. Then, the difference between the paired response scores from BioRank 2009 and 2018 were calculated. The response scores were separated by body region, and the mean difference for each body region was calculated. Additionally, the standard deviation of the paired differences was calculated using the following equation:

$$S_d = \sqrt{\frac{\sum_{i=1}^{p}(d_i - d_m)^2}{p-1}} \quad (1)$$

where S_d is the standard deviation of the paired differences, d_i is the difference between paired values, d_m is the mean of the differences within a body region, and p is the number of responses in a given body region. Using the equation for the t-statistic for paired data, the critical difference was calculated as follows:

$$d_0 = d_m - t\frac{S_d}{p} \quad (2)$$

where d_0 is the critical difference for a body region, t is the t-statistic for α=0.05 and p-1 degrees of freedom, and d_m, S_d, and p are as defined above for Equation 1. Finally, the critical differences were averaged across body regions to calculate an overall critical difference that could be compared to the difference between overall scores from BioRank 2009 and 2018 for each ATD. A difference in BioRank 2009 and 2018 scores greater than the critical difference would indicate a significant difference between the two versions.

The final analysis was to evaluate the degree of correlation between the biofidelity response scores from BioRank 2009 and 2018. Because the data did not follow a normal distribution, the Spearman rank correlation, a non-parametric test, was used to evaluate whether correlations between biofidelity scores were statistically significant. The correlations of interest included comparing the 2009 and 2018 biofidelity response scores, the difference in the 2009 and 2018 scores versus the 2009 scores, and the SM and P scores from BioRank 2018 with the 2018 biofidelity response scores (RMS). The difference between the 2009 and 2018 scores was compared to

the 2009 scores to assess whether there was a relationship between the version differences and the magnitude of the scores, i.e., were differences between BioRank versions larger for higher scores. Additionally, the shape and magnitude (SM) and phase (P) scores were compared to the BioRank 2018 RMS scores to evaluate which component had a greater effect on the RMS score. Despite the skewed distributions of the data, it was still of interest to assess whether the relationships between the parameters of interest were approximately linear. Therefore, a linear regression was applied to each comparison and the slope, y-intercept, and R^2 values were calculated.

III. RESULTS

The BioRank 2018 response scores were generally lower than the BioRank 2009 response scores, meaning BioRank 2018 reported greater biofidelity for both ATDs compared to BioRank 2009. For the HIII, 14 response scores increased from version 2009 to 2018, while 87 scores decreased. For the THOR, 22 scores increased and 79 scores decreased. Measures of central tendency and other parameters describing the differences between the 2009 and 2018 signal response scores are provided in Table I for each ATD. The hierarchical averaging scheme resulted in overall biofidelity scores of 2.42 and 2.15 for BioRank 2009 and 2018, respectively, for the HIII. For the THOR, the overall scores were 3.24 and 3.11 for BioRank 2009 and 2018, respectively.

TABLE I
MEASURES OF CENTRAL TENDENCY FOR THE DIFFERENCE BETWEEN BIORANK 2018 AND BIORANK 2009 RESPONSE SCORES (DIFFERENCE = 2018 - 2009)

Measure	HIII	THOR
Mean	-0.26	-0.17
Absolute Mean	0.31	0.37
Median	-0.18	-0.15
Absolute Median	0.21	0.20
Mean Increase	0.18	0.45
Mean Decrease	-0.33	-0.35
Maximum Increase	0.45	6.30
Maximum Decrease	-1.43	-2.30

The response scores between BioRank 2009 and 2018 were statistically different; however, the overall biofidelity scores were not. Specifically, the Wilcoxon Signed Rank test showed that the individual biofidelity response scores from BioRank 2009 and 2018 were significantly different ($p<0.0001$) for both ATDs. The critical differences and version differences for each ATD are compared in Figure 2. The critical differences varied across body region, particularly for the THOR. For most body regions, the critical difference was greater than the version difference, except for the HIII thorax. This trend carried through into the overall average, indicating that remaining version differences after the averaging scheme were not significant. The THOR had very large scores for the knee, which resulted in large differences and a large critical difference. Therefore, the overall average was also calculated while excluding the knee response as a potential outlier.

All correlations of interest between BioRank response scores were statistically significant ($p<0.0001$). Table II shows the Spearman correlation coefficient (ρ) and R^2 value for each correlation. The biofidelity response scores from BioRank 2009 and 2018 had a strong, positive, and linear correlation for both ATDs (Figure 3). The response scores from BioRank 2009 had a strong negative correlation with the difference between version scores (2018-2009), when using the non-parametric model. In other words, as the magnitude of the scores increased, the decrease from BioRank 2009 to 2018 increased as well. This was reflected in a modest linear correlation in the HIII (Figure 4), but little linear correlation in the THOR (Figure 5). However, the highest scoring outliers from the knee excursions seemed to be skewing the linear regression. Removing the knee excursions from the THOR analysis resulted in a modest linear correlation as well (Figure 6). Converting the differences between versions to absolute differences slightly increased the R^2 values for the linear regressions for both the HIII and THOR (with and without knees excluded). When comparing the P and SM components to the RMS (response score) for

BioRank 2018, SM showed a high linear correlation with RMS (Figure 7). While P had a significant positive correlation with RMS according to the non-parametric test, P had a much weaker linear correlation with RMS than SM. The R^2 values for P were about 1/3 of that of SM, indicating that SM was a greater contributor to the RMS score than P.

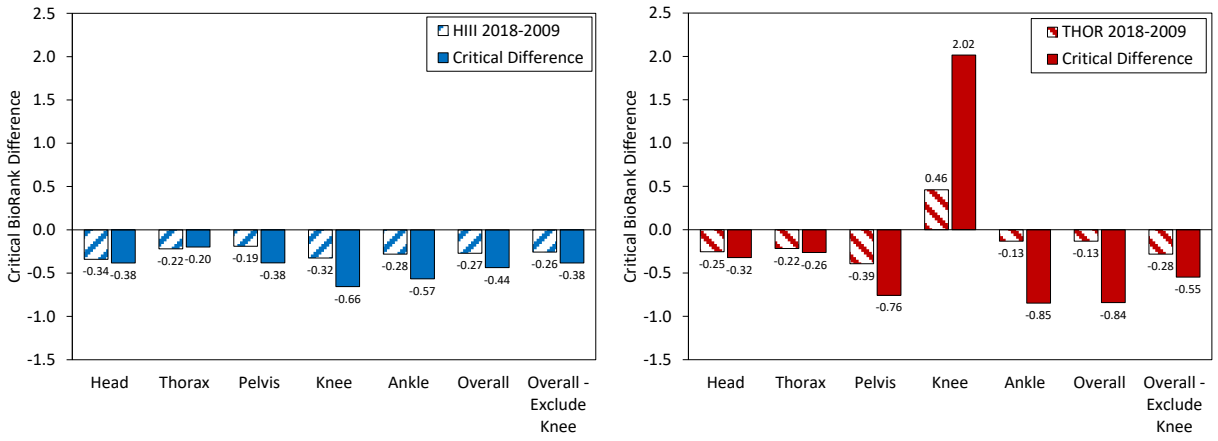

Fig. 2. Version differences in BioRank scores and critical differences for the HIII (left) and THOR (right) for all body regions and overall.

TABLE II
CORRELATION PARAMETERS

Independent Variable	Dependent Variable	ρ		R^2	
		HIII	THOR	HIII	THOR
BioRank 2009	*BioRank 2018*	0.98	0.97	0.96	0.94
BioRank 2009	*BioRank (2018-2009)*	-0.52	-0.38	0.30	0.05
BioRank 2009	*BioRank |2018-2009|*	0.66	0.70	0.44	0.28
BioRank 2018 SM	*BioRank 2018 RMS*	0.99	0.99	0.98	1.00
BioRank 2018 P	*BioRank 2018 RMS*	0.57	0.62	0.35	0.28

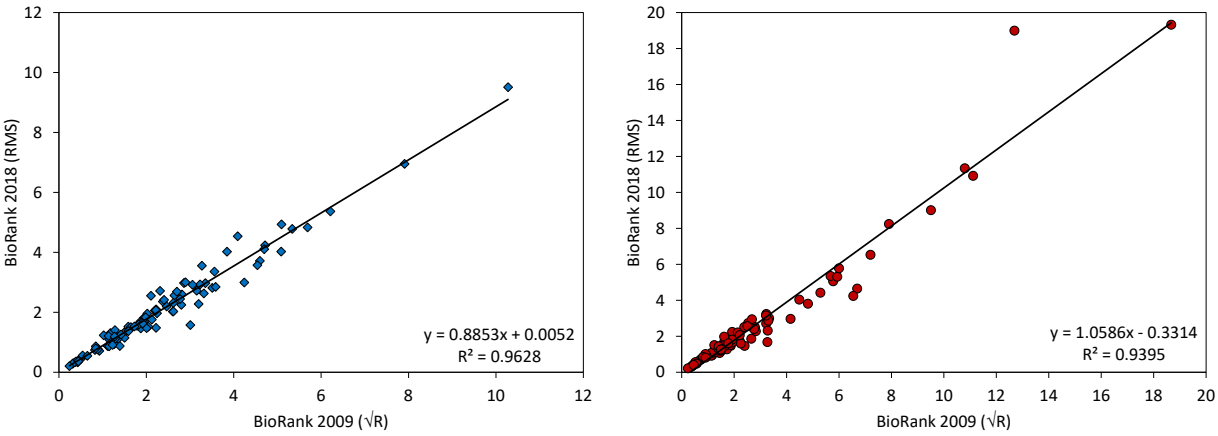

Fig. 3. Correlations between BioRank 2009 and BioRank 2018 response scores for the HIII (left) and THOR (right).

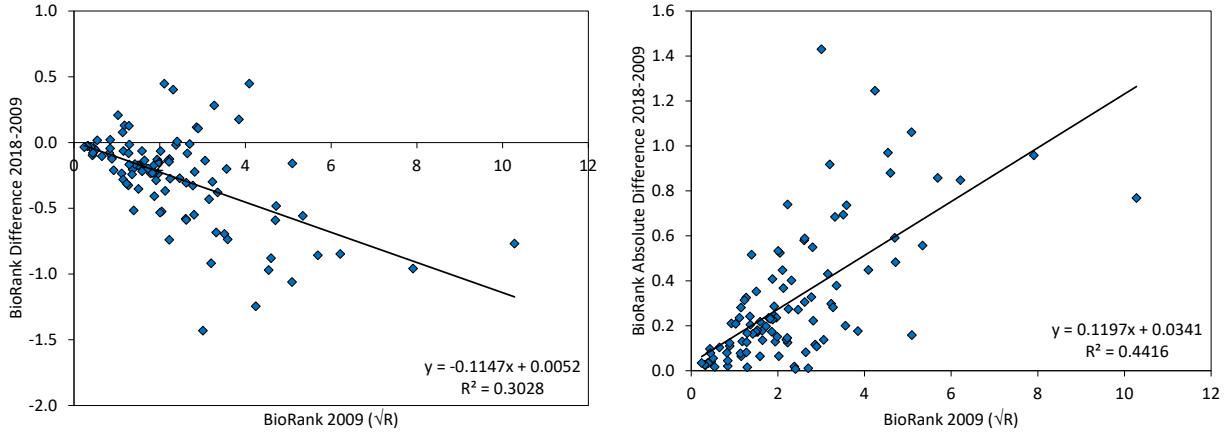

Fig. 4. Correlations between BioRank 2009 and the signed (left) and absolute (right) differences between BioRank 2018 and 2009 response scores for the HIII.

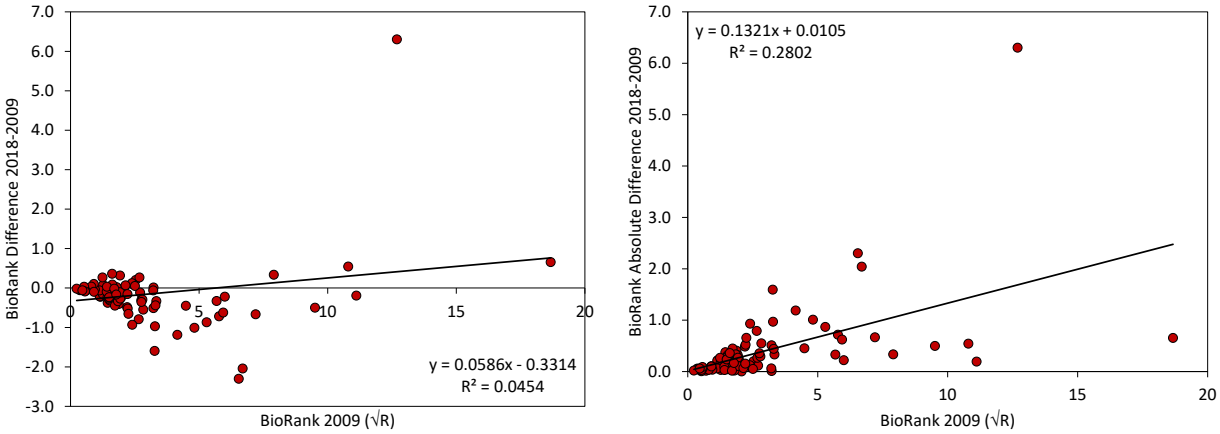

Fig. 5. Correlations between BioRank 2009 and the signed (left) and absolute (right) differences between BioRank 2018 and 2009 response scores for the THOR.

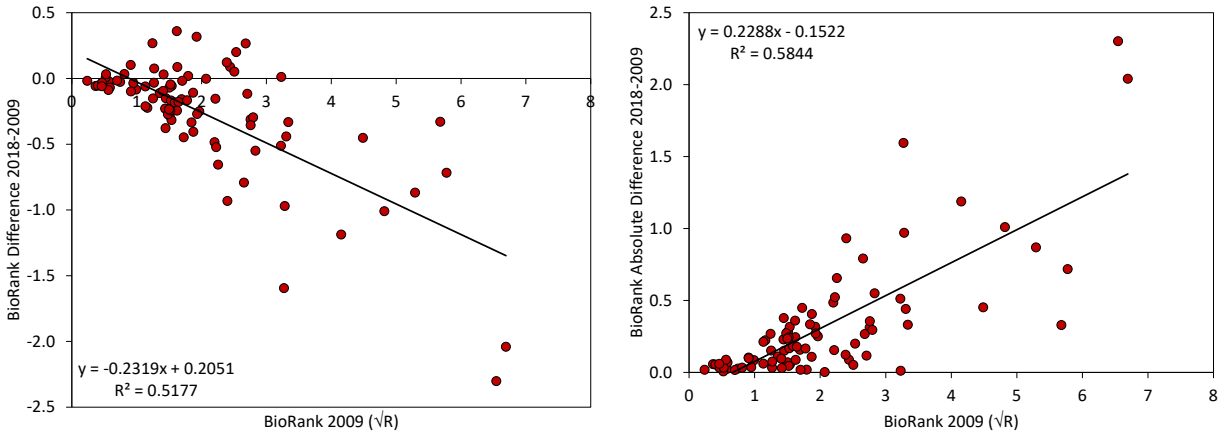

Fig. 6. Correlations between BioRank 2009 and the signed (left) and absolute (right) differences between BioRank 2018 and 2009 response scores for the THOR with the knee responses excluded.

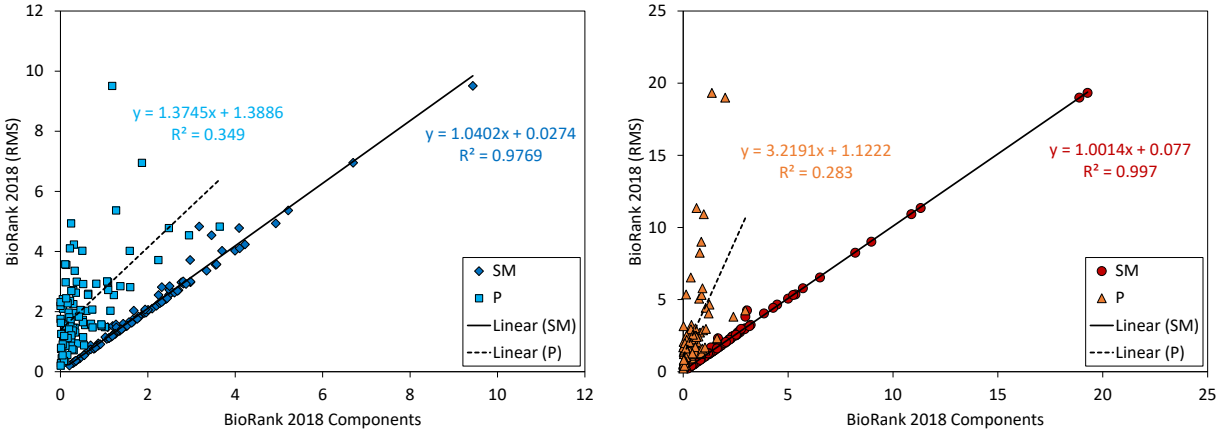

Fig. 7. Correlations between BioRank 2018 RMS and BioRank 2018 components, SM and P, for the HIII (left) and THOR (right).

IV. Discussion

The application of both BioRank 2009 and 2018 to a dataset consisting of matched HIII and THOR frontal sled tests demonstrated that the two versions of BioRank yielded significant score differences depending on the response level of interest. On the level of individual response measurements, BioRank 2018 produced significantly lower biofidelity scores than BioRank 2009. However, once responses were averaged into specific body regions and into an overall score using a hierarchical averaging scheme, differences between versions were no longer statistically significant. This finding has different implications on studies in the literature depending on how a specific study used BioRank. As noted in the Introduction, some studies do not implement any type of averaging, but only report the biofidelity scores of individual response signals [7][8][13-17][19][20]. These studies are more likely to be affected by version differences than those that use a hierarchical averaging structure because individual response scores were found to be statistically different between BioRank versions, while hierarchically averaged scores were not.

Two other deviations from the hierarchical averaging used in BioRank have been observed in the literature. The first deviation observed in three studies in the literature is a straight (non-hierarchical) average where all biofidelity response scores are averaged together regardless of test condition, body region, and signal type [9][11][12]. The second deviation from the traditional averaging scheme can be found in the study that evaluated the biofidelity of the THOR-50M [3]. The study combined a large number of component tests on different body regions as well as sled tests performed under four different test conditions to assess the biofidelity of the THOR-50M compared to the HIII. For all of the component tests, the hierarchical averaging scheme from BioRank 2009 was followed. However, the sled tests were grouped into a separate group the study termed *whole body*, which was treated like an additional body region in the averaging scheme. Consequently, responses in the sled tests were averaged within a test condition without being first separated into different body regions. Then the test condition averages were averaged to create a whole body score, which was then averaged with the other body region scores from the component tests to produce overall biofidelity scores. Since the dataset used in the current study is from sled tests, the whole body averaging approach and the straight average approach were implemented to assess whether different averaging approaches could lead to non-trivial differences in the overall biofidelity scores. Table III shows the results of this analysis. Both the whole body and straight average approaches result in lower (better) biofidelity scores compared to the hierarchical average for both ATDs. In fact, the THOR would have poor biofidelity according to the hierarchical averaging method, but marginal biofidelity according to the other two methods (see BioRank scale classification in [3][5]). Additionally, the critical difference analysis was repeated using the straight average and whole body averaging schemes to create analogous critical difference values. The results of the analysis is shown in Figure 8. The difference between the overall biofidelity scores for BioRank 2009 and 2018 surpass the critical difference for both the straight average and whole body average when the knee is excluded from the critical difference benchmark for the THOR. This indicates that the differences between overall biofidelity scores between BioRank 2009 and 2018 are significant under the straight and whole body averaging schemes. Therefore, the averaging scheme used in a particular study may contribute to whether using a different version of BioRank will influence the findings of that study.

TABLE III
OVERALL BIOFIDELITY SCORES USING DIFFERENT AVERAGING SCHEMES

Averaging	HIII		THOR	
Scheme	2009	2018	2009	2018
Straight	2.32	2.06	2.72	2.55
Whole Body	2.32	2.06	2.73	2.55
Hierarchical	2.42	2.15	3.24	3.11

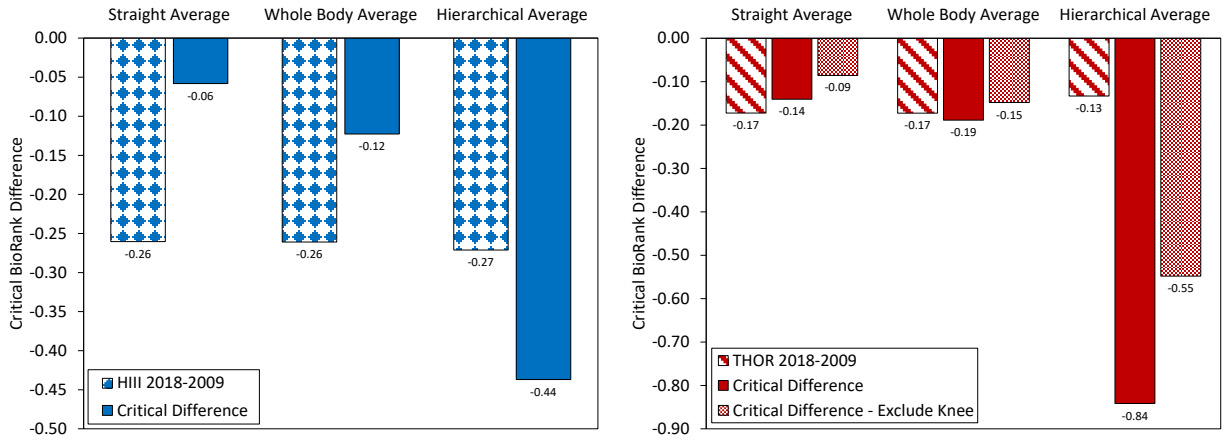

Fig. 8. The difference between overall BioRank scores (2018-2009) and the associated critical differences using three averaging schemes for the HIII (left) and THOR (right).

Another factor that could influence the BioRank scores in a study is the type of human response corridor used for the BioRank calculation. As discussed previously, there are several methods that can be used for calculating a standard deviation corridor, and sometimes a standard deviation corridor is not available in the literature for a particular dataset. To demonstrate how different standard deviation calculations can influence BioRank results, both BioRank 2009 and 2018 were recalculated using sample standard deviation corridors instead of population standard deviation corridors. The overall biofidelity scores that resulted from both definitions of standard deviation, using hierarchical averaging, are provided in Table IV. Using sample standard deviation drastically reduced the biofidelity scores for both BioRank versions as a result of increased corridor width for all responses. Notably, the HIII shifted from having marginal biofideltiy to good biofidelity, while THOR shifted from having poor biofideltiy to marginal biofidelity. This is a relatively extreme example due to the low sample size used to construct the corridors in this study (2-3 PMHS), and differences between population and sample standard deviation will decrease as sample sizes increase. However, this brief analysis demonstrates the importance of controlling for and being transparent about the type of standard deviation or corridor used in BioRank. Ideally, if a non-standard corridor formulation is used, a sensitivity analysis should be conducted to evaluate the effect of changes to corridor width on BioRank scores.

TABLE IV
OVERALL BIOFIDELITY SCORES USING DIFFERENT DEFINITIONS OF STANDARD DEVIATION

Standard	HIII		THOR	
Deviation Type	2009	2018	2009	2018
Population	2.42	2.15	3.24	3.11
Sample	1.85	1.63	2.46	2.33

How internal and external responses are defined and averaged within the BioRank averaging scheme also has the potential to influence the overall BioRank scores and the conclusions of biofidelity analyses. For example, the biofidelity study of the THOR-50M described above deviated from the definitions of internal and external responses used in BioRank 2009, particularly by classifying body region excursions as external [3]. This classification was more similar to the internal and external definitions provided in BioRank 2002. To evaluate how

this difference can affect overall biofidelity scores, the BioRank 2002 definitions of internal and external signals were applied to the dataset in the current study using the hierarchical and whole body averaging schemes. Interestingly, applying the BioRank 2002 definitions to the hierarchical analysis decreased overall biofidelity scores for both ATDs and both versions of BioRank (2009 and 2018), while applying the 2002 definitions to the whole body analysis increased scores (Table V). Additionally, repeating the whole body analysis on the data from [3] using the BioRank 2009 definitions of internal and external yielded different conclusions for that particular *body region* compared to the original study. The THOR internal biofidelity score increased from 1.472 to 1.619 and its external score increased from 1.989 to 2.532. The HIII scores also increased with the internal score increasing from 1.576 to 1.622, and the external score increasing from 1.780 to 2.075. Finally, the overall biofidelity scores for the whole body region (the average of the internal and external scores), which were not reported in the original study, increased from 1.730 to 2.075 for the THOR and increased from 1.678 to 1.849 for the HIII. Using the categorical classification system, the HIII retained its good biofidelity rating while the THOR dropped to a rating of marginal.

TABLE V
OVERALL BIOFIDELITY SCORES WITH DIFFERENT DEFINITIONS OF INTERNAL AND EXTERNAL RESPONSES

Internal/External Reference	Averaging Scheme	HIII		THOR	
		2009	2018	2009	2018
BioRank 2002	Hierarchy	2.38	2.08	2.85	2.64
BioRank 2009	Hierarchy	2.42	2.15	3.24	3.11
BioRank 2002	Whole Body	2.39	2.11	2.93	2.75
BioRank 2009	Whole Body	2.32	2.06	2.73	2.55

In all of the alternate analyses conducted in the Discussion so far, the overall biofidelity score for BioRank 2018 has consistently been lower than that of BioRank 2009 for both ATDs. Two methodological differences between BioRank versions may have led to the pervasive decrease in BioRank 2018 scores. First, the method for calculating the shape and magnitude component of the score was changed from using cumulative variance (a summation of squared differences) in BioRank 2009 to a summation of absolute differences in BioRank 2018. Essentially, this changed the function that computed and penalized differences between response curves from a quadratic function to a linear function. Accordingly, these functions will perform differently, alternately elevating and depressing scores relative to each other depending on whether score differences are greater than or less than one. Specifically, the cumulative variance (quadratic) method will produce higher scores (a higher penalty), compared to the absolute difference method (linear), when differences between response curves are greater than one. Conversely, the absolute difference method will impose a higher penalty for differences between response curves that are less than one. As a result, ATD curves that are very similar to the human response curve will have slightly increased scores for BioRank 2018, while ATD curves with a component outside of one standard deviation will experience score decreases using BioRank 2018. In other words, BioRank 2018 may reduce the quantitative biofidelity rating of already excellent responses, but will improve the biofidelity rating of good, marginal, and poor responses. Figure A2 in the Appendix illustrates both of these cases using simulated human and ATD response curves. In this example, a relatively a small part of ATD1's response curve is outside of the human response corridor, yielding a BioRank 2009 score of 1.11 (good biofidelity) and a better score of 0.86 (excellent biofidelity) for the SM component of BioRank 2018. Conversely, ATD2's response curve is very similar to the human response curve, but its score slightly increases from 0.16 to 0.18 when changing from BioRank 2009 to 2018. A limitation of this example is the assumption that BioRank 2009 and BioRank 2018 evaluate response differences under the same phase conditions, i.e., neither version uses phase optimisation prior to the calculation. This assumption may not be valid in all cases due to the phase optimisation procedure in BioRank 2018, where the ATD response curve is phase shifted to generate the best possible RMS score. This optimisation is the second factor that may contribute to the decreased scores associated with BioRank 2018.

It is difficult to evaluate the relative contributions of the phase optimisation procedure and computational differences between BioRank versions to the score reductions in BioRank 2018. Comparing the BioRank 2018 SM scores to the BioRank 2009 response scores contains the contributions of both mechanisms. However, calculating

the 2018 SM score without phase shifting the ATD curve isolates the effect of using the cumulative absolute difference mechanism. When unshifted versions of the 2018 SM scores were compared to the BioRank 2009 response scores, the cumulative absolute difference method produced a decrease in 76 response measurements for the HIII and 71 response measurements for the THOR out of a total of 101 for each ATD. The median change was -0.12 and -0.06 for the HII and THOR, respectively. The mean change was -0.10 for the HIII and -0.02 for the THOR. Therefore, for the curves in this dataset, the cumulative absolute difference method alone produced decreases, albeit small, in scores for both ATDs. However, it appears that both mechanisms are actively contributing to the score decreases observed at the response level since the average differences are much greater between the BioRank 2009 and 2018 biofidelity response scores.

The division of the biofidelity response score into phase (P) and shape and magnitude (SM) components was mostly driven by the discovery that BioRank 2002 and 2009 failed to produce subjectively accurate scores for short duration, high magnitude responses when the ATD and human response curves were out of phase [2][27]. The intent of the score division was to permit out of phase responses to be aligned for the SM computation, increasing the SM score. However, this increase in the SM score was balanced by introducing the P score, which acted as a penalty if a phase shift was necessary. It was of interest to determine whether the SM and P scores were contributing equally to the magnitude of the RMS score so a correlation analysis between the BioRank RMS and its score components was performed. The results indicate that the SM score is contributing more than the P score to the RMS. This is visually evident in the higher degree of correlation between the RMS and SM scores compared to the RMS and P scores (Figure 7). The SM and RMS correlations also had R^2 values that were approximately three times higher than the P and RMS correlation (Table II), indicating that the SM score is contributing more to the variance in the RMS score. The differences in contribution seem to be a direct result of the magnitude disparity between the SM and P scores. The median SM scores were 1.63 and 1.51 for the HIII and THOR, respectively. Conversely, the median P scores were 0.26 and 0.24 for the HIII and THOR, respectively. Because these scores are equally weighted in the RMS calculation, but the P score is generally much lower than the SM score, the SM score will have a greater contribution to the resulting RMS score. The P score was derived to represent the number of standard deviations the ATD curve is away from the mean human response curve via a phase shift. As a result, an ATD curve would have to be at least one standard deviation of phase shift away from the human response to generate a score comparable to the median SM score. It would be uncommon for ATD or model responses to be largely out of phase compared to the human response. Hence, lower values of P compared to SM are likely the norm in a typical dataset to which BioRank is applied. Therefore, the RMS score may be implicitly biased toward SM over P. Whether this discrepancy in relative contribution between the two components improves or diminishes the accuracy of BioRank is difficult to assess and would require a larger analysis with a wider breadth of signal types, which is outside of the scope of this study. However, a case study evaluating the effect of the P score on a high magnitude and short duration signal is included in the Appendix.

This study is not intended to be used as an extrapolation tool to convert from one version of BioRank to another. Since every dataset has different limitations and results, only using and comparing both versions on a dataset of interest can quantify how using different versions could alter the biofidelity scores of that dataset. The dataset used in the current study is limited in several ways. First, only 2-3 PMHS were used to build the mean PMHS responses and corridors. Second, only three whole body test conditions (no component tests) were used, and those test conditions were very similar with the only difference being what restraints were used. Third, the analysis was only performed on two 50^{th} percentile male frontal ATDs evaluated in a full-frontal crash mode. If this type of analysis was performed using data from other types of ATDs, in terms of demographic representation and intended crash mode, results and conclusions may vary. Fourth, only internal response measures were used in this study. Lastly, no effort was made to balance the number of responses between body regions so certain body regions could have a greater effect on the average biofidelity scores than others.

Despite its limitations, this study shows that using BioRank 2018 instead of BioRank 2009 can significantly reduce biofidelity scores in a statistical sense. Therefore, it is likely that applying BioRank 2018 to the THOR-50M biofidelity analysis would lead to reduced (better) biofideltiy scores. Likewise, applying BioRank 2009 to the THOR-05F biofidelity analysis may lead to increased (worse) biofidelity scores. In the current study, the HIII seemed to benefit from a greater reduction in overall biofidelity scores compared to the THOR through the use of BioRank 2018. Since both the HIII and THOR-50M were evaluated and compared in [3], it would be interesting to apply BioRank 2018 to the data in that study for both ATDs and observe whether their comparative biofidelities

remain the same. In the current study, the overall HIII biofidelity scores were always lower (better) then the THOR scores regardless of how the analysis was performed. However, this may not be the case with all datasets. Considering a biofidelity analysis using a different objective rating metric on the data in this study also found the HIII to be slightly more biofidelic than the THOR-50M [31][32], the result may be implicit to this particular dataset.

It is important to note that a statistically significant difference is not always a practically meaningful difference. An example of this can be seen in Figure 2, where the critical (statistically significant) difference for the THOR knee is over 2.0. However, BioRank scores are intended to represent standard deviations away from a mean. Hence, this result implies that only differences greater than two standard deviations are meaningful when comparing BioRank scores, which is not logical given that categorical differences in biofidelity scores are defined in increments of one standard deviation. It is often difficult to quantify the concept of a meaningful difference. Although the categorical differences between biofidelity scores are defined in increments of one, it seems that differences less than one could be meaningful as well, especially if they happen to bridge categories. Perhaps relating differences in biofidelity scores to differences in injury outcomes would provide some meaning to incremental differences. Unfortunately, such an analysis is outside of the scope of this paper. Additionally, it is difficult to quantify how differences between BioRank versions could impact the overall conclusions of studies that use BioRank because it is unknown which version of BioRank provides a more accurate assessment of biofidelity. Given that BioRank was developed to be used as a standalone measure of biofidelity for a single ATD or model, it is important to assess the relative accuracies of different BioRank versions, which provide different biofidelity scores. However, the best method for evaluating the accuracy of different versions of BioRank is currently unclear, and such an analysis is outside of the scope of this study.

The author would like to propose several recommendations regarding the use of BioRank based on the results of this study. First, it is important to clearly report which version of BioRank is being used for an analysis and if there are any deviations in the calculations, averaging scheme, etc., from how a particular version was originally defined. In particular, it is important to report how the PMHS corridors were calculated, i.e., what method was used to calculate standard deviation, and whether/how the response data were truncated to a specific interval of evaluation. Furthermore, if non-standard corridors must be used, it would be appropriate to perform and report a sensitivity analysis regarding how varying corridor width affects the resulting BioRank scores. Next, different definitions of internal and external responses and whether/how these responses are differentiated when averaging across responses can influence the overall biofidelity score(s) and assessment. Therefore, it should be explicitly reported what responses in a study are considered internal and external, and definitions should be kept consistent across body regions and test conditions. If a hierarchical averaging scheme is incorporated that differentiates internal and external responses, care should be taken that the number of external and internal responses across different body regions is relatively balanced. Performing the analysis with and without separating the external and internal responses could provide an indication as to whether an imbalance is skewing results. Similarly, the type of averaging scheme used (if any) should be carefully considered to prevent too much or too little weight from being given to a particular body region, particularly if the number of signals varies between body regions. Finally, critical differences from a critical difference analysis should only be applied to the particular dataset from which they were produced since variation between (and within) datasets can produce variations in critical differences.

V. CONCLUSIONS

This study showed that BioRank 2018 resulted in significantly lower biofidelity response scores than BioRank 2009 for this particular dataset. Additionally, the overall biofidelity scores could be significantly different between BioRank versions, depending on the type of averaging scheme used. These results indicate that studies using one version of BioRank may obtain statistically different results if another BioRank version were used instead. As an example, the THOR-50M biofideltiy evaluation was performed using BioRank 2009 so using BioRank 2018 may significantly decrease (improve) the biofidelity scores in that evaluation. Conversely, the biofidelity of THOR-05F was assessed using BioRank 2018; therefore, performing the analysis with BioRank 2009 could lead to increased (worse) biofidelity scores. However, it is currently unclear what quantity constitutes a meaningful difference in biofidelity scores between different versions of BioRank and which BioRank version is providing a more accurate assessment of biofidelity. Therefore, more work is needed before the implications of this study can be fully understood.

VI. ACKNOWLEDGEMENT

The author would like to thank Dr. Andrew Kemper for providing valuable feedback on this manuscript and the study contained herein.

VII. REFERENCES

[1] Rhule, H.H., Maltese, M.R., et al. Development of a new biofidelity ranking system for anthropomorphic test devices. *SAE Technical Paper*, 2002

[2] Moorhouse, K., Donnelly, B., Kang, Y.-S., Bolte IV, J.H., and Herriott, R. Evaluation of the internal and external biofidelity of current rear impact ATDs to response targets developed from moderate-speed rear impacts of PMHS. *SAE Technical Paper*, 2012

[3] Parent, D., Craig, M., and Moorhouse, K. Biofidelity Evaluation of the THOR and Hybrid III 50th Percentile Male Frontal Impact Anthropomorphic Test Devices. *Stapp Car Crash Journal*, 2017. 61: p. 227-276.

[4] Stammen, J., Moorhouse, K., Suntay, B., Carlson, M., and Kang, Y.-S. The large omnidirectional child (LODC) ATD: biofidelity comparison with the hybrid III 10 year old. *SAE Technical Paper*, 2016

[5] Wang, Z.J., Lee, E., et al. Biofidelity Evaluation of THOR 5th Percentile Female ATD. *Proceedings of IRCOBI Conference*, 2018. Athens, Greece.

[6] Yaek, J., Li, Y., et al. Biofidelity assessment of the 6-year-old ATDs in lateral impact. *Traffic Injury Prevention*, 2016. 17(5): p. 535-543.

[7] Asanuma, H. and Takahashi, Y. Improvement and Validation of the Lower Limb and the Pelvis for a Pedestrian Dummy. *SAE Technical Paper*, 2015

[8] Bose, D., Subit, D.L., et al. Biofidelity improvements to the Polar-II pedestrian dummy lower extremity. *SAE Technical Paper*, 2007

[9] Kang, Y.-S., Moorhouse, K., Herriott, R., and Bolte IV, J.H. Comparison of cervical vertebrae rotations for PMHS and BioRID II in rear impacts. *Traffic Injury Prevention*, 2013. 14(sup1): p. S136-S147.

[10] Konosu, A., Issiki, T., and Tanahashi, M. Development of a biofidelic flexible pedestrian leg-form impactor (Flex-PLI 2004) and evaluation of its biofidelity at the component level and at the assembly level. *SAE Transactions*, 2005: p. 2298-2311.

[11] Parent, D., Craig, M., Ridella, S., and McFadden, J.D. Thoracic biofidelity assessment of the THOR Mod Kit ATD. *Proceedings of 23rd Enhanced Safety of Vehicles Conference*, 2013. Seoul, Republic of Korea.

[12] Ramachandra, R., Kang, Y., Hagedorn, A., Stammen, J., and Bolte IV, J. Abdominal Biofidelity Assessment of 50th Percentile Male and 10-Year-Old ATD Responses Relative to a Recently Developed Belt-Loading Corridor. *Proceedings of IRCOBI Conference*, 2017. Antwerp, Belgium.

[13] Stammen, J.A., Donnelly, B.R., Suntay, B., and Moorhouse, K.M. Dynamic Response Criteria for a Large Child ATD Thoracic Spine. *Proceedings of IRCOBI Conference*, 2014. Berlin, Germany.

[14] Takahashi, Y., Kikuchi, Y., et al. Biofidelity evaluation for the knee and leg of the polar pedestrian dummy. *Proceedings of International Technical Conference on the Enhanced Safety of Vehicles*, 2005. Washington D.C., USA.

[15] Antona-Makoshi, J., Yamamoto, Y., et al. Age-dependent factors affecting thoracic response: a finite element study focused on Japanese elderly occupants. *Traffic Injury Prevention*, 2015. 16(sup1): p. S66-S74.

[16] Asanuma, H., Takahashi, Y., Ikeda, M., and Yanaoka, T. Investigation of a Simplified Vehicle Model that Can Reproduce Car-Pedestrian Collisions. *SAE Technical Paper*, 2014

[17] Dokko, Y., Yanaoka, T., and Ohashi, K. Validation of age-specific human FE models for lateral impact. *SAE Technical Paper*, 2013

[18] Golman, A.J., Danelson, K.A., Gaewsky, J.P., and Stitzel, J.D. Implementation and validation of thoracic side impact injury prediction metrics in a human body model. *Computer Methods in Biomechanics and Biomedical Engineering*, 2015. 18(10): p. 1044-1055.

[19] Ito, Y., Dokko, Y., Motozawa, Y., Mori, F., and Ohashi, K. Kinematics Validation of Age-Specific Restrained 50 th Percentile Occupant FE Model in Frontal Impact. *SAE Technical Paper*, 2012

[20] Ramachandra, R., Kang, Y., et al. Evaluation of Skeletal and Soft Tissue Contributions to Thoracic Response of GHBMC M50-O Model in Dynamic Frontal Loading Scenarios. *Proceedings of IRCOBI Conference*, 2019. Florence, Italy.

[21] Vavalle, N.A., Jelen, B.C., Moreno, D.P., Stitzel, J.D., and Gayzik, F.S. An evaluation of objective rating methods for full-body finite element model comparison to PMHS tests. *Traffic Injury Prevention*, 2013. 14(sup1): p. S87-S94.
[22] Vavalle, N.A., Moreno, D.P., Rhyne, A.C., Stitzel, J.D., and Gayzik, F.S. Lateral impact validation of a geometrically accurate full body finite element model for blunt injury prediction. *Annals of Biomedical Engineering*, 2013. 41(3): p. 497-512.
[23] Wang, Y., Kim, T., Li, Y., and Crandall, J. Neck Validation of Multibody Human Model under Frontal and Lateral Impacts using an Optimization Technique. *SAE Technical Paper*, 2015
[24] Thunert, C. CORA Release 3.6 User's Manual. 2012, GNS mbH, Germany.
[25] ISO. Road vehicles - Objective rating metric for non-ambiguous signals. 2014, ISO.
[26] Rhule, H. Improvements to NHTSA's Biofidelity Ranking System and Application to the Evaluation of the THOR 5th Female Dummy. *Proceedings of IRCOBI Conference*, 2018. Athens, Greece.
[27] Rhule, H., Donnelly, B., Moorhouse, K., and Kang, Y.S. A methodology for generating objective targets for quantitatively assessing the biofidelity of crash test dummies. *Proceedings of Enhanced Safety of Vehicles*, 2013. Seoul, Republic of Korea.
[28] Rhule, H., Moorhouse, K., Donnelly, B., and Stricklin, J. Comparison of WorldSID and ES-2re biofidelity using an updated biofidelity ranking system. *Proceedings of International Technical Conference on the Enhanced Safety of Vehicles*, 2009. Stuttgart, Germany.
[29] Kang, Y.-S., Bolte IV, J.H., et al. Biomechanical responses of PMHS in moderate-speed rear impacts and development of response targets for evaluating the internal and external biofidelity of ATDs. *SAE Technical Paper*, 2012
[30] Irwin, A.L., Sutterfield, A., et al. Side impact response corridors for the rigid flat-wall and offset-wall side impact tests of NHTSA using the ISO method of corridor development. *SAE Technical Paper*, 2005
[31] Albert, D.L., Beeman, S.M., and Kemper, A.R. Occupant kinematics of the Hybrid III, THOR-M, and postmortem human surrogates under various restraint conditions in full-scale frontal sled tests. *Traffic Injury Prevention*, 2018. 19(sup1): p. S50-S58.
[32] Albert, D.L., Beeman, S.M., and Kemper, A.R. Assessment of thoracic response and injury risk using the Hybrid III, THOR-M, and post-mortem human surrogates under various restraint conditions in full-scale frontal sled tests. *Stapp Car Crash Journal*, 2018. 62: p. 1-65.
[33] SAE. Instrumentation for Impact Test - Part 1 - Electronic Instrumentation. 2014, SAE International.

VIII. APPENDIX

BioRank Version History

This section provides a detailed summary of all versions of BioRank. A tabular summary of the following information is provided in Table AI.

BioRank 2002 introduced a two part system that first assessed the biofidelity of individual ATD response measurements, and then combined those measurements via an averaging scheme to assess overall ATD biofidelity. In order to assess the biofidelity of individual ATD response time histories, BioRank 2002 divided the cumulative variance between the mean ATD and human response curves (DCV) by the cumulative variance between the mean curve and one standard deviation curve of the human response (CCV). The biofidelity score ($\sqrt{R}$) assigned to each response measurement was then the square root of this ratio. The score was intended to represent how many standard deviations the ATD response was from a typical human response. For example, a score less than two would indicate that the ATD response falls within two standard deviations of a human response. Therefore, a lower BioRank score indicated better biofidelity. After calculating the biofidelity score for each response measurement, a hierarchical averaging and weighting scheme was implemented to combine the scores of all response measurements into a single overall biofidelity score for the ATD. First, the responses were divided into internal and external responses. According to BioRank 2002, internal responses are responses that are measured within the ATD and have an associated potential injury criterion. External responses are made external to the ATD and describe how the ATD interacts with the vehicle components in a crash. For example, head acceleration, head displacement measured via video analysis or motion capture, and a pendulum force from an impact to the head would be classified as internal, external, and external responses, respectively. Throughout the hierarchical averaging scheme, internal and external responses were evaluated separately until the last step. First, all internal and external responses were separately averaged within a test condition for each body region. Here, the ranking system assumed that multiple types of tests had been conducted on an ATD before biofidelity was assessed. Therefore, test conditions could consist of sled tests at different speeds or in different configurations, component-level tests on different body regions, or any combination thereof and more. Then, a weighted average of each test condition was calculated, such that the result was an internal and external score for each body region. The weights were intended to be determined by the user for their particular application and dataset with the guidance that 33% of the weight should come from a *subject score* and 67% of the weight should come from a *test relevance score*. The subject score ranged from one to 10 and was assigned based on how many human subjects or PMHS were used to develop the response corridors within a particular test condition. The test relevance score also ranged from one to 10 and was assigned based on how well the ATD responses in a particular condition matched the human response targets in regulatory-type tests. After the weighted average of test conditions was computed, the scores of all body regions were averaged to generate overall internal and external biofidelity scores. As a final step, the internal and external biofidelity scores could be averaged to produce a final overall biofidelity score. See Figure 1 in the Methods section for a schematic representation of the hierarchical averaging scheme for internal responses.

The second version of BioRank (BioRank 2009) made no changes to how the individual response scores were calculated, but made several changes to the hierarchical averaging scheme [28]. First, the test condition weights were removed. Therefore, to generate the biofidelity scores at the body region level, an unweighted average was calculated using all test condition scores within a particular body region. Responses were still separated by their internal and external classifications until the end of the averaging scheme. However, the internal and external classification system was revised. This was stated in the text as a loosening of the definition of the internal response to include internal measures without associated injury criteria. In addition, externally measured displacements, such as head displacement, were reclassified as internal responses instead of external responses. Detailed definitions of internal and external responses within the framework of BioRank 2009 can be found in [29].

The third version of BioRank (BioRank 2013) made recommendations for data processing prior to calculating BioRank and modified the calculation of the response scores [27]. Data processing recommendations included procedures for normalising and performing phase optimisation on human response curves as needed prior to generating the mean human response and standard deviation corridors. Additionally, the generation of constant width standard deviation corridors, using the average standard deviation across time for a response, was

recommended to eliminate the necking that occurs when the human response curves happen to be very similar or intersect. It was erroneously stated that this would not affect the calculation of the cumulative variance between the mean and one standard deviation human response curves [26]. In fact, this does alter the cumulative variance calculation because the differences between the mean and standard deviation curves are squared before being summed. It was also suggested that the calculation of response scores be limited to the portion of the mean human response curve that was within 80% of the peak response, minimising the inclusion of low magnitude data outside of the event of interest. Next, the calculation of the response biofidelity score was divided into two parts: a phase score (P) and a shape and magnitude score (SM). The SM score was calculated using the same methodology that was used to calculate √R in BioRank 2002 and BioRank 2009, with the exception that the ATD response was first phase optimised with respect to the mean human response curve via cross-correlation. However, there are several methods of calculating cross-correlation and a specific method or equation was not provided. The phase score was calculated as the shift needed to phase optimise the ATD response for the SM calculation divided by the *acceptable lag*. The acceptable lag was generated by shifting the mean human response curve with respect to itself until the √R value calculated between the curves reached 1.0. The root mean square (RMS) of SM and P was defined as the new biofidelity response score.

The fourth version of BioRank (BioRank 2018) modified the calculation of the response biofidelity score and proposed a modification to the hierarchical averaging scheme [26]. The calculation of the SM score was modified to be the cumulative absolute difference between the ATD and mean human response curves (DCAD) divided by the cumulative absolute difference between the mean and one standard deviation human response curves (CCSD). Essentially, this modified the penalty function for differences between curves from being second order (a squared term) in BioRanks 2002, 2009, and 2013 to first order (an absolute difference) in BioRank 2018. Additionally, it allowed the application of constant standard deviation corridors as described in BioRank 2013 without altering the calculation of the biofidelity score. The phase optimisation procedure for the ATD response prior to the calculation of SM was also altered. Instead of a cross-correlation function, the ATD phase shift was defined as the shift that produced the minimum biofidelity response score. The P score was calculated in a similar method as described in BioRank 2013. P was still the ratio between the ATD phase shift and the acceptable lag. However, the acceptable lag was generated by shifting the mean human response curve with respect to itself until the cumulative absolute difference between the shifted and unshifted curves exceeded the CCSD. During the shifting procedure, the human response curve was be padded to maintain a constant interval of evaluation for the cumulative absolute differences. As in BioRank 2013, an RMS calculation was used to combine the SM and P scores into the response biofidelity score. For the hierarchical averaging changes, it was stated that it is not necessary to evaluate internal and external biofidelity separately, as in previous versions of BioRank, and that they could be grouped together within test condition averages.

TABLE AI

SUMMARY OF DIFFERENCES BETWEEN BIORANK VERSIONS

BioRank Version	Pre-processing	Response Score Calculation	Averaging Scheme
2002		• Uses cumulative variance	• Separates internal and external responses • Applies test condition weights when averaging
2009		• Uses cumulative variance	• Separates internal and external responses • Discontinues use of test condition weights
2013	• Suggests normalisation and phase optimisation of human responses before averaging into corridors • Suggests generating constant width standard deviation corridors	• Uses cumulative variance • Separates response score into sub-scores for shape and magnitude (SM) and phase (P) • Phase optimises ATD response using cross-correlation	• Separates internal and external responses
2018	• Suggests generating constant width standard deviation corridors	• Uses cumulative absolute difference instead of cumulative variance • Continues use of separate SM and P scores • Phase optimises ATD response by finding the phase shift that generates the lowest response score	• Suggests separating internal and external responses is not necessary

BioRank Calculations

The biofidelity response scores for BioRank 2009 were calculated using the following set of equations [28]:

$$DCV = \sum_{t=0}^{n}[D(t) - C_m(t)]^2 \quad \text{(A1)}$$

$$CCV = \sum_{t=0}^{n}[C_{SD}(t) - C_m(t)]^2 \quad \text{(A2)}$$

$$\sqrt{R} = \sqrt{\frac{DCV}{CCV}} \quad \text{(A3)}$$

where n is the length of the response curve, DCV is the ATD cumulative variance, D(t) is the ATD response curve, $C_m(t)$ is the mean PMHS response curve, CCV is the PMHS cumulative variance, $C_{SD}(t)$ is the upper PMHS standard deviation curve, and $\sqrt{R}$ is the biofidelity response score. BioRank 2009 was calculated without performing phase optimisation on the ATD response curve with respect to the mean PMHS response curve since phase optimisation was not introduced until BioRank 2013.

In order to calculate biofidelity using BioRank 2018, a number of steps, including phase optimisation of the ATD response curve relative to the mean PMHS response curve, were performed. Determination of the phase optimised ATD response is governed by the allowable phase shift. In order to determine the allowable phase shift, the PMHS cumulative standard deviation was first calculated using the following equation:

$$CCSD = \sum_{t=0}^{n}|C_{SD}(t) - C_m(t)| \quad \text{(A4)}$$

where CCSD is the PMHS cumulative standard deviation and $C_{SD}(t)$ and $C_m(t)$ are defined as above for Equations A1-A2. Next, the mean PMHS response curve was shifted left and right relative to itself until the cumulative standard deviation of the shifted curve relative to the unshifted curve was greater than the CCSD. Essentially, the minimum left and right shifts were found such that the following inequality was satisfied:

$$CCSD < \sum_{t=0}^{n}|C_s(t \pm b) - C_m(t)| \quad \text{(A5)}$$

where $C_s(t)$ is the shifted mean PMHS response curve and b is the set of time steps by which the curve is shifted. The minimum positive shift was defined as the allowable right shift, and the minimum negative shift was defined as the allowable left shift. Next, Equations A6-A9 were calculated for each possible left and right shift of the ATD response curve relative to the mean PMHS response curve. When, the ATD curve was shifted, the left or right side of the PMHS mean response curve was padded, as applicable, to maintain a constant interval of evaluation for Equations 6-9. Note that the possible phase shift is limited to the length of the PMHS response curve and can be greater than the allowable left and right shifts calculated using Equation A5.

$$DCAD_{\pm b} = \sum_{t=0}^{n}|D(t \pm b) - C_m(t)| \quad \text{(A6)}$$

$$SM_{\pm b} = \frac{DCAD_{\pm b}}{CCSD} \quad \text{(A7)}$$

$$P_{\pm b} = \begin{cases} \frac{b}{APS_R}, & b > 0 \\ \frac{b}{APS_L}, & b < 0 \\ 0, & b = 0 \end{cases} \quad \text{(A8)}$$

$$RMS_{\pm b} = \sqrt{{SM_{\pm b}}^2 + {P_{\pm b}}^2} \quad \text{(A9)}$$

DCAD is the ATD cumulative difference and b is the set of left and right time shifts of the ATD response curve. SM is the shape and magnitude score. P is the phase score. APS_L and APS_R are the left and right allowable phase shifts respectively. RMS is the root mean square of the SM and P scores. The shifted ATD response curve that resulted in the lowest value of RMS was considered the phase optimised ATD response curve. Therefore, the minimum RMS that resulted from all possible phase shifts was the biofidelity response score for BioRank 2018.

The same hierarchical averaging scheme was used for both BioRank 2009 and 2018 in the current study. An equation representation of the averaging scheme used in the current study is as follows:

$$B = \frac{\sum_{i=1}^{l}\left(\frac{\sum_{j=1}^{m}\left(\frac{\sum_{k=1}^{p}\left(\sqrt{R_{i,j,k}}\ or\ RMS_{i,j,k}\right)}{p}\right)}{m}\right)}{l} \quad \text{(A10)}$$

where B is the overall biofidelity (or BioRank) score, l is the number of body regions, m is the number of test conditions within a body region, p is the number of response measurements within a test condition and body region, √R is the BioRank 2009 biofidelity score for a particular response measurement, and RMS is the BioRank 2018 biofidelity score for a particular response measurement. Table AII shows all of the response measurements included in this study divided by body region.

TABLE AII
ALL RESPONSE MEASUREMENTS DIVIDED INTO BODY REGIONS

Body Region	Response Measurement	Conditions Included
Head	CG Acceleration X	All
	CG Acceleration Z	All
	CG Angular Velocity X	All
	CG Angular Velocity Y	All
	CG Angular Velocity Z	All
	CG Excursion X	All
	CG Excursion Z	All
Thorax	Chest Acceleration X	All
	Chest Acceleration Z	All
	Chest Angular Velocity X	All
	Chest Angular Velocity Y	All
	Chest Angular Velocity Z	All
	Left Shoulder Excursion X	All
	Left Shoulder Excursion Z	All
	Right Shoulder Excursion X	All
	Right Shoulder Excursion Z	All
	Upper Sternum Deflection	All
	Upper Left Chest Deflection	All
	Upper Right Chest Deflection	All
	Lower Left Chest Deflection	All
	Lower Right Chest Deflection	All
Pelvis	Pelvis Acceleration X	All
	Pelvis Acceleration Z	All
	Pelvis Angular Velocity X	All
	Pelvis Angular Velocity Y	All
	Pelvis Angular Velocity Z	KB/SWAB, KBAB/SWAB
Knee	Left Knee Excursion X	All
	Left Knee Excursion Z	All
	Right Knee Excursion X	All
	Right Knee Excursion Z	All
Ankle	Left Ankle Excursion X	All
	Left Ankle Excursion Z	All
	Right Ankle Excursion X	All
	Right Ankle Excursion Z	All

Case Study of High Magnitude, Short Duration Signals

A case study was undertaken within the confines of the current study to evaluate whether the addition of the phase score improves the subjective accuracy of BioRank for high magnitude, short duration signals. The best example of such a signal within the current study is the forward head acceleration during the KB condition, where the head strikes the steering wheel/hub for all surrogates. The characteristic average responses for each ATD are shown in Figure A1 along with the mean PMHS response with one standard deviation corridors, the BioRank 2009 score (√R), and the BioRank 2018 SM, P, and RMS scores. The HIII response was a higher magnitude and out of phase compared to the PMHS response. This resulted in a very poor BioRank 2009 score over 6.0. The 2018 P score of 1.27 accurately captured that the HIII response was out of phase, and the SM score showed approximately a 1.0 reduction in score compared to BioRank 2009. Overall, the phase optimisation in BioRank 2018 resulted in a lower response score compared to BioRank 2009 for the HIII, but the difference in magnitude between the responses prevented a substantial improvement in the BioRank score even when the phasing difference was considered. Conversely, the THOR BioRank scores for the same response increased from the 2009 version to the 2018 version. Since the THOR signal seemed only slightly out of phase, the increase seemed to be driven by an increase of the SM score that was independent of the phase correction. Instead, the change from using the cumulative variance in BioRank 2009 to the cumulative absolute difference in BioRank 2018 appears to have caused the increase. This limited analysis demonstrates that the phase shift can provide some correction to improve BioRank scores for short duration, high magnitude, out of phase signals; however, this correction can be dwarfed by the contribution of the SM component.

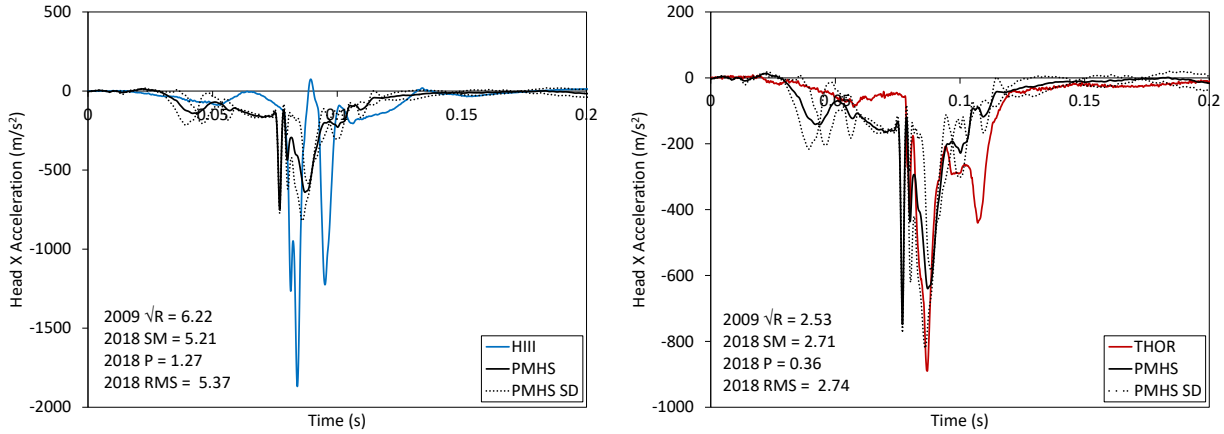

Fig. A1. Forward head accelerations of the HIII (left) and THOR (right) compared to the PMHS response for the KB condition.

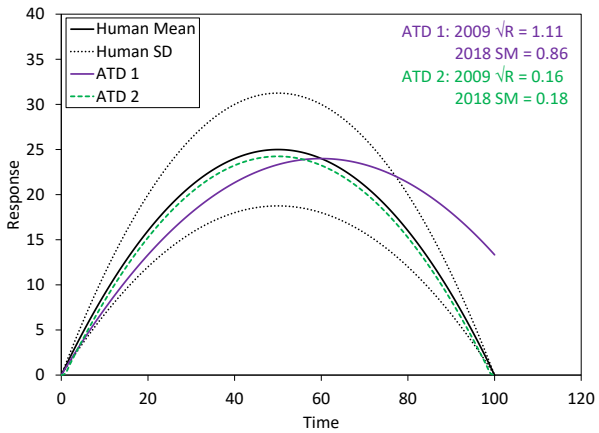

Fig. A2. Simulated human and ATD response curves illustrating differences between the 2009 BioRank score (2009 √R) and non-phase optimised 2018 BioRank shape and magnitude score (2018 SM).