Three Approaches to Biofidelity Evaluation: A Comparative Study

Jeremie Peres

Abstract Biofidelity evaluation is essential for assessing the accuracy of human surrogates in crash safety research. This study compares three biofidelity assessment methods — BioRank, CORA, and the novel BioScore, to investigate their alignment, strengths, and limitations. A dataset of post-mortem human subject (PMHS) and dummy response signals was analysed, with biofidelity scores calculated using each method. To facilitate cross-method score interpretation, a unified grading scale was developed. The findings reveal that the choice of assessment method significantly affects the biofidelity outcomes, with each method exhibiting distinct limitations. Notably, BioRank tends to underestimate signals with accurate shape but poor amplitude and overestimate those with correct amplitude but poor shape. BioScore may overestimate scores in presence of high PMHS variability, while CORA struggles when the average PMHS curve does not represent accurately the shape of the individual PMHS curves. Despite these limitations, both CORA and BioScore show promise as reliable biofidelity evaluation tools, given targeted refinements and further validation. This study underscores the need for ongoing improvements and standardisation in biofidelity assessment practices to enhance surrogate accuracy and drive vehicle safety advancements.

Keywords Biofidelity, BioRank, BioScore, CORA, crash safety, curve rating.

I. INTRODUCTION

As the development of human surrogates, such as crash test dummies and virtual human body models (HBMs), advances, the need for accurate biofidelity comparisons becomes increasingly critical. Several methods have been proposed for biofidelity evaluation, yet no consensus has been reached on the best approach. For crash test dummies, the BioRank method [1] developed by the National Highway Traffic Safety Administration (NHTSA), is widely used, whereas for HBMs, variations of the CORA method [2-3] are commonly employed. However, there are few direct comparisons between these two methods that highlight their respective strengths and limitations.

This paper presents a comparison of three biofidelity evaluation methods. The objective is to determine whether the BioRank and CORA methods produce significantly different outcomes. Both methods involve comparing the dummy curve to the average post-mortem human subject (PMHS) data, which may introduce issues when the curves exhibit differences in shape or phase. To mitigate this, a third method (BioScore) was introduced, which compares the dummy curve to each individual PMHS curve rather than relying on the mean.

II. METHODS

This study compares three biofidelity evaluation methods: BioRank, CORA, and BioScore. The initial step involved testing each method individually to identify the most effective implementation. Subsequently, the correlations between the three methods using regression analysis were assessed. BioRank includes a predefined grading scale that classifies numerical values from *marginal* to *excellent*. To facilitate a more comprehensive comparison, not only based on numerical values but also on qualitative interpretation, corresponding scales for CORA and BioScore were developed. The following paragraphs describe the dataset used, the BioRank, CORA, and BioScore methods, along with the approach used to define the grading scales for CORA and BioScore.

J. Peres (e-mail: jeremie.peres@pdb-org.com, tel: +49 (0) 1525 7710736) is a Research Engineer at the Partnership for Dummy Technology and Biomechanics (PDB), Gaimersheim, Germany.

Test Dataset

Reference [4] evaluated the biofidelity of the H350 and THOR-50M dummies using the BioRank method. The biofdelity assessment was performed by comparing PMHS mean curves and corridors from various component and whole-body tests to paired dummy test data. A significant portion of the original PMHS and dummy data is available in the National Highway Traffic Safety Adiministration (NHTSA) biomechanics database [5].

For this study, a subset of the load cases was selected, primarily based on the availability of individual PMHS curves and the corresponding dummy curves. Three load cases were chosen, Gold Standard 2, Gold Standard 3, and Far-Side Oblique, resulting in a total of 85 curve comparisons: 43 for the THOR-50M dummy and 42 for the H350. All curves were time-aligned by identifying the time zero in the pulse, resampled at 10kHz and filtered according to the ISO 6487:2015 [6]. For each signal, an average curve and corridors representing ±1 and ±2 standard deviations (SD) were generated using time-averaging. It is important to note that [4] used a different implementation of the BioRank method, and the paper does not specify the preprocessing steps applied to the curves. As a result, the BioRank scores calculated in this study may differ from those reported in the original study.

Biofidelity Evaluation Methods

BioRank

In 2002, NHTSA introduced a methodology to objectively assess dummy biofidelity [1]. This method has since been progressively refined and updated [7-10]. It provides a comprehensive assessment framework, from establishing biofidelity targets to aggregating individual signal scores into an overall biofidelity score.

Although successive versions of BioRank employ different objective curve rating schemes, the core principle remains consistent: the dummy response is compared to the average PMHS response within ±1 standard deviation corridors. The initial implementation calculated the ratio of the cumulative variance between the mean PMHS curve and the dummy curve, relative to the cumulative variance between the mean PMHS curve and the upper bound of the SD corridor.

In later versions [9], this ratio, termed the shape and magnitude rating (SM_t), replaced cumulative variance with the cumulative absolute difference ($CCSD$ & $DCAD$). This adjustment provided a more intuitive representation of the deviation between the dummy and PMHS curves, expressing the difference as a multiple of the PMHS data's standard deviation. Additionally, a phase rating (P_t) was introduced, and the BioRank score (RMS_t) was calculated as the Euclidean norm of SM_t and P_t. The phase shift t was determined by minimising the BioRank value. More details on the calculation of CCSD and DCAD are provided in Appendix 1.

$$RMS_t = \sqrt{SM_t^2 + P_t^2}, \text{ (1)}$$

with:

$$SM_t = DCAD/CCSD, \text{ (2)}$$

$$P_t = \frac{DummyPhaseShift}{AcceptablePhaseShift}, \text{ (3)}$$

The most recent update to BioRank, published in 2022 [10], introduced two key changes. First, the phase shift was obtained by optimising SM_t rather than RMS_t. Second, the primary score is reduced to SM_t, while P_t and RMS_t are still calculated and documented as supplementary information. This study focuses on the last two versions of BioRank.

BioRank scores are expressed as multiples of the standard deviation, for example a score of 1 indicates that the dummy curve is one SD away from the mean PMHS curve. A scale was defined for BioRank for qualitative interpretation of the results [4][8]:

TABLE I
BIORANK GRADING SCALE

Score	Grade
B ≤ 1	Excellent
1 < B ≤ 2	Good
2 < B ≤ 3	Fair
3 < B	Marginal

CORA

In 2009, [2] introduced an objective rating method that combines a corridor score with a cross-correlation score to evaluate the accuracy of simulation models against the corresponding dummy test results. The cross-correlation method assesses phase shift, magnitude, and shape, providing a score between 0 (no correlation) and 1 (perfect correlation) for each aspect. The overall rating is calculated as a weighted average of the individual scores. Later, the ISO TC22/SC10/SC12/WG4 group proposed an enhanced metric [11], ISO 18571, integrating the corridor rating from CORA with the phase, magnitude, and slope ratings from the Enhanced Error Assessment of Response Time Histories (EEARTH) [12] (see Appendix 1 for more details), as illustrated below:

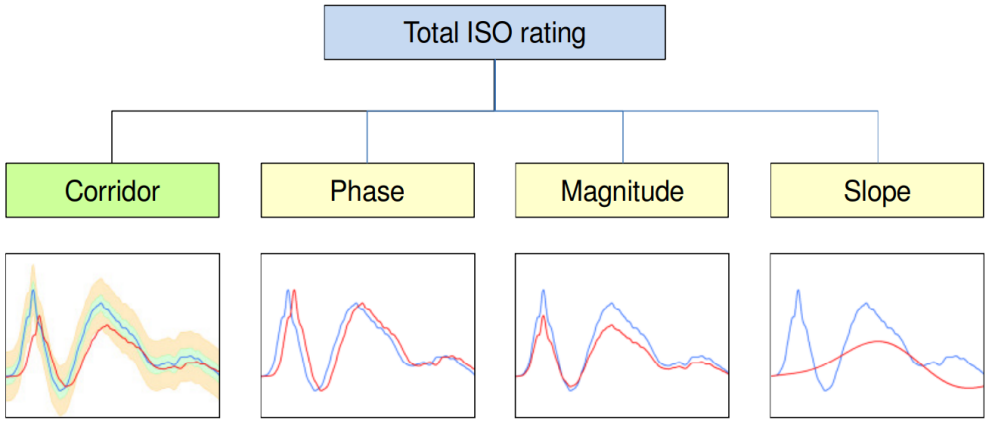


Fig. 1. Structure of the ISO metric *[3]*. Green rectangle: CORA, Yellow rectangles: EEARTH. Blue curve: Reference, Red curve: Simulation.

As with CORA, this metric mitigates the limitations of individual ratings by combining multiple measures. The slope score improves on the shortcomings of CORA and other methods by using a dynamic time-warping algorithm which reduces the influence of time discrepancies. In 2013, ISO 18571 [11] was evaluated [3] on representative examples and showed that it *can handle different kinds of non-ambiguous signals of different qualities*.

In the corridor method, inner and outer corridors are defined around the reference curve. By default, these corridors are generated automatically, with the inner corridor width set usually to 5% and the outer corridor width set usually to 50% of the reference signal's absolute maximum amplitude. Alternatively, the CORA method allows the definition of user-defined corridors instead of the automatically generated ones.

Most studies evaluating the biofidelity of HBMs have used CORA as an objective rating method ([13-30]). In the majority of cases, the corridors were user-defined, with the inner and outer ones set as multiples of ±1 SD.

In the current study, the phase, magnitude, and slope ratings from ISO 18571 were used, along with two variants of inner and outer corridor definitions: Variant 1 defined both the inner and outer corridors as ±1 standard deviation (SD), while Variant 2 defined the inner corridor as ±1 SD and the outer corridor as ±2 SD.

BioScore.

BioRank and CORA utilise the comparison of the dummy or HBM curve with the average of the PMHS curves. However, it is well established that averaging can result in a mean curve that does not accurately represent the shape or characteristics of any individual curve, especially when there are significant phase and shape differences among them [31-32]. In 2007, [33] proposed a new method that compares the dummy curve directly to each individual PMHS curve, rather than to a mean curve and corridors. The individual comparison is performed using a cross-correlation function that evaluates phase, magnitude, and shape. This method is considered a potential alternative to BioRank and CORA. To take advantage of the advancements and standardisation efforts in curve comparison since the method's publication, CORA was configured to use the metric of ISO 18571 for the individual curve comparisons.

The method, illustrated in Figure 2, is based on calculating a Dummy-to-PMHS score, which represents the average score obtained by comparing each PMHS curve to the dummy curve. Additionally, a PMHS-to-PMHS score is calculated as the average score between individual PMHS curves. The PMHS-to-PMHS score reflects the

variability within the PMHS data, while the Dummy-to-PMHS score indicates how well the dummy curve aligns, on average, with the PMHS data. The final BioScore is computed as the ratio of the Dummy-to-PMHS score to the PMHS-to-PMHS score:

$$BioScore = \frac{Dummy\ To\ PMHS\ score}{PMHS\ To\ PMHS\ score}, (4).$$

Two variants of the BioScore method were computed, differing in how the Dummy-to-PMHS score is calculated. When calculating the CORA score between two signals, one is defined as the reference curve and the other as the simulation curve. CORA is not commutative, meaning that the score differs when the reference and simulation curves are exchanged. In the first variant, all possible combinations between dummy and PMHS curves are included, even those where the dummy curve serves as the reference. In the second variant, combinations in which the dummy curve is defined as the reference are excluded:

Figure 2: Illustration of the Dummy (red)-to-PMHS (yellow) score and PMHS-to-PMHS score [33].

Development of Grading Scales for CORA and BioScore

To the author's knowledge, no grading scale currently exists for CORA when used for biofidelity evaluation, nor for BioScore. Defining a scale for CORA and BioScore would facilitate comparisons between the three tested methods, making it easier to interpret the results beyond purely numerical values. To achieve this, the BioRank scale was used as a reference, trying to define corresponding thresholds for CORA and BioScore. It can be defined as a classification problem [34] in which the best match in classes for the curves tested between the three methods is to be found. This was achieved by moving the thresholds in a predefined range and finding the thresholds values which lead to the best values for F1-Score. F1-Score is a metric which evaluates the performance of a classification model, it is better suited than accuracy for unbalanced datasets [35].

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}, (5).$$

with:

$$Precision = \sum_i^N \frac{True\ Positives\ i}{True\ Positives\ i + False\ Positives\ i}, (6).$$

and:

$$Recall = \sum_i^N \frac{True\ Positives\ i}{True\ Positives\ i + False\ Negatives\ i}, (7).$$

where i is a specific class and N is the total number of classes. A true positive is a case where the model correctly predicts a positive outcome, a false positive is a case where the model incorrectly predicts a positive outcome when it should have been negative, and a false negative is a case where the model incorrectly predicts a negative outcome when it should have been positive.

Three strategies were tested to align the grading scales: (1) optimizing the BioScore scale against BioRank, then aligning CORA to BioScore; (2) optimizing the CORA scale against BioRank, then aligning BioScore to CORA; and (3) jointly optimizing both CORA and BioScore scales against BioRank in parallel.

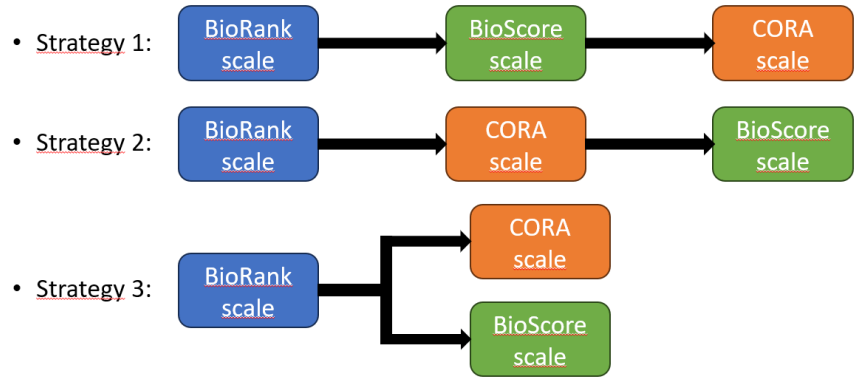


Fig. 3. Illustration of the three strategies pursued to determine CORA and BioScore grading scales.

Strategies 1 and 2 may seem less intuitive, but they are expected to improve agreement between CORA and BioScore and could potentially lead to overall better F1-scores across the three metrics. Once the grading scales for CORA and BioScore were defined, dummy signals were assigned a grade for each method.

III. RESULTS

Implementations of BioRank and CORA

Generally, the scores produced by the 2018 and 2022 versions of BioRank correlate strongly (R^2 = 0.94) as illustrated in Figure 5. BioRank 2022 systematically yields higher scores due to the removal of the phase component, which can lead to significant phase shifts for certain signals without affecting the score (see Figure 4). However, large phase adjustments appear unjustified when pulse alignment is already achieved, as residual phase differences likely reflect genuine mechanical behaviour disparities. While the 2018 version of BioRank may not always perfectly align signal characteristics, it at least accounts for phase differences, with any residual misalignment ultimately reflecting inherent differences between the dummy and PMHS. Therefore, the 2018 implementation of the BioRank method was chosen for the remainder of the study.

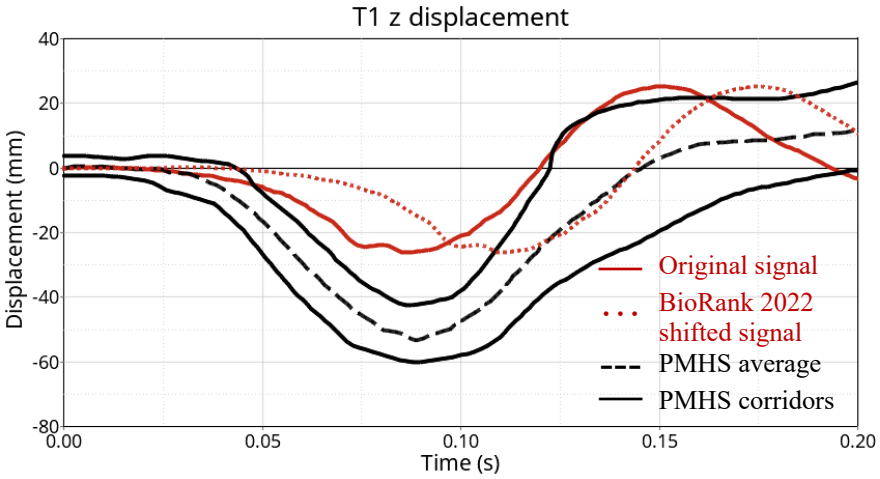


Fig. 4. Example of large phase shift with BioRank 2022

The two implementations of the CORA method, based on different definitions of the corridors yielded highly similar results (R^2 = 0.98, slope = 1.01, see Figure 6). Given that the corridor rating accounts for only 40% of the overall score in this study, this similarity is not entirely surprising. To maintain consistency with BioRank, which uses a single set of corridors corresponding to ±1 SD, variant 1 was selected, which applies 1 SD for both the inner and outer corridors.

Although the two implementations of the BioScore algorithm correlated strongly (R^2 = 0.95), the values differed, as indicated by the slope and intercept of the regression line (Figure 7). For the individual curve-to-curve scores, the corridor width is determined as a percentage of the maximum absolute amplitude of the reference curve. When the dummy curve is used as the reference, there is a risk that an artificial peak could result in

unrealistically wide corridors, even though this scenario was not observed in the present study. Additionally, it is more logical to use PMHS curves as the reference, since the goal is to compare the dummy to the PMHS, not the reverse. Given these considerations, variant 2 of BioScore was selected.

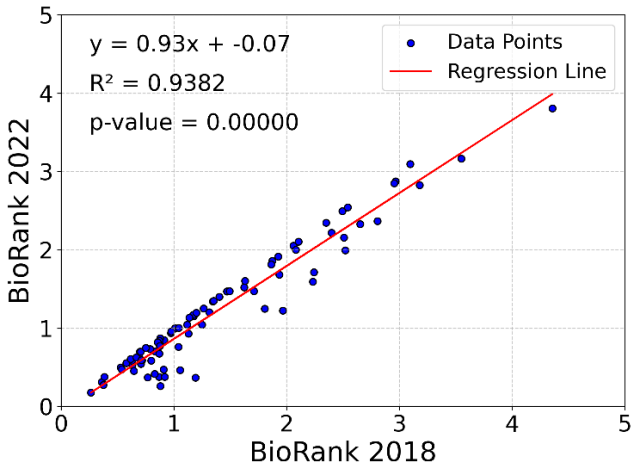


Fig. 5. Regression between the 2018 and 2022 versions of BioRank.

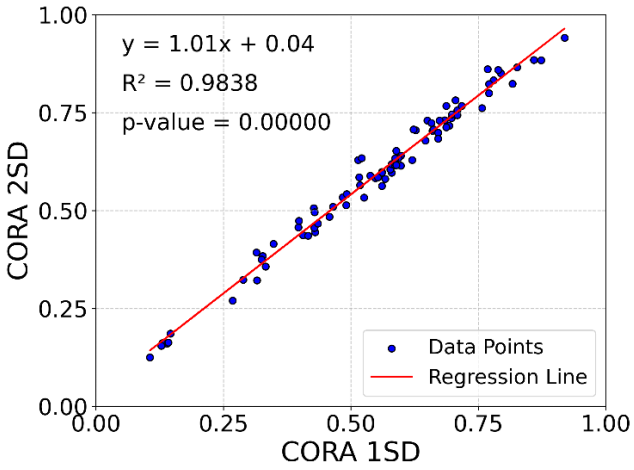


Fig. 6. Regression between the CORA with variant 1 and variant 2 corridors.

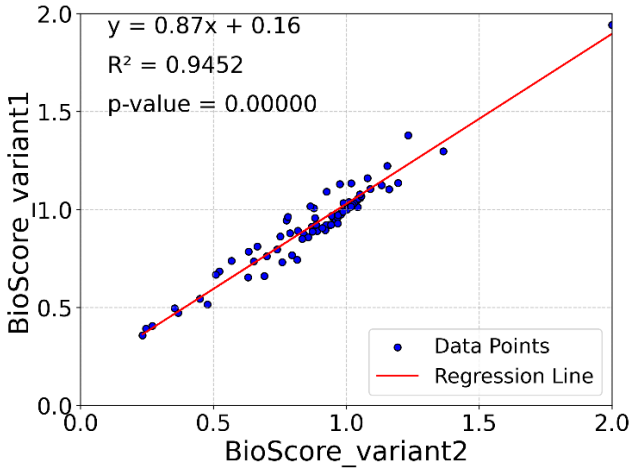


Fig. 7. Regression between variant 1 and 2 of BioScore.

Correlation Between the Scores of the Three Methods

The p-values associated with the regressions between the three methods are all below 0.05 (see Figure 8, Figure 9, and Figure 10), indicating statistically significant relationships. However, the R-squared values are low, suggesting that these relationships explain only a relatively small portion of the variability. When examining the relationship between CORA and BioScore, most data points appear to correlate well, but the R-squared value seems to be negatively impacted by a few outliers, where BioScore values are disproportionately high compared to CORA. A closer examination of these data points reveals that they correspond to cases with limited available PMHS curves, which also exhibit substantial variability (PMHS-to-PMHS score < 0.5). In contrast, BioRank generally shows poor correlation with the other two methods. Overall, these results indicate that the choice of method for performing a biofidelity evaluation has a significant impact on the outcomes.

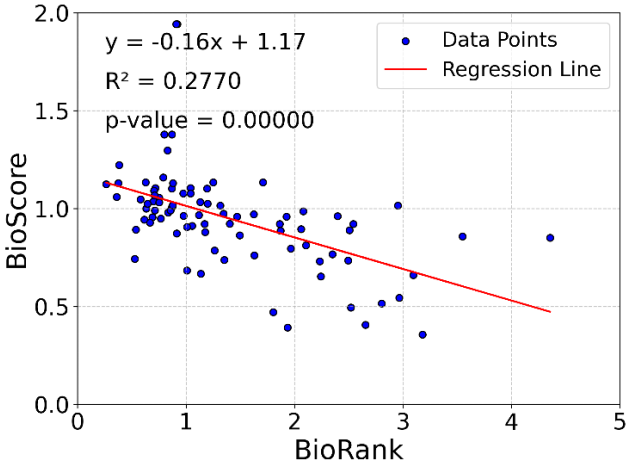


Fig. 8. Regression between BioRank and BioScore.

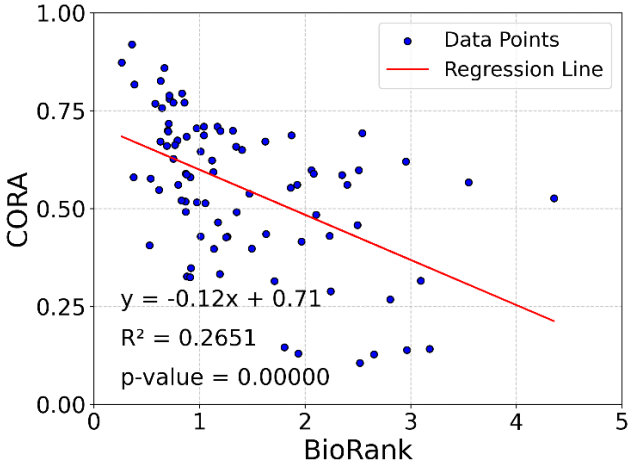


Fig. 9. Regression between BioRank and CORA.

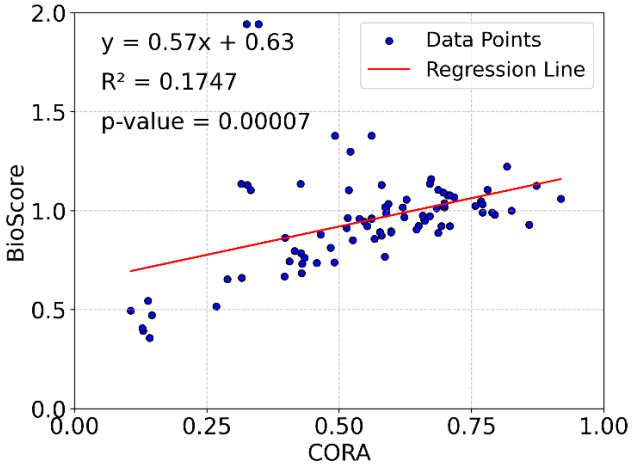


Fig. 10. Regression between CORA and BioScore.

Determination of the Grading Scales for CORA and BioScore

A very small number of curves (n = 4) received a marginal grade (B > 3) according to the BioRank grading scale, making it difficult to establish a threshold between fair and marginal. Therefore, these two classes were combined into a single category called *Fair or Marginal*. Figure 11 and Figure 12 illustrate the distribution of all data points for CORA and BioScore, with each point coloured according to the BioRank scale. The task of determining the grading scales for the two methods reduces to identifying two horizontal lines that separate the points into the categories *Fair or Marginal*, *Good*, and *Excellent*. However, since there are no clear separations between these categories, whatever thresholds is selected, some points classified into one category by BioRank will be assigned to a different category by the other two methods.

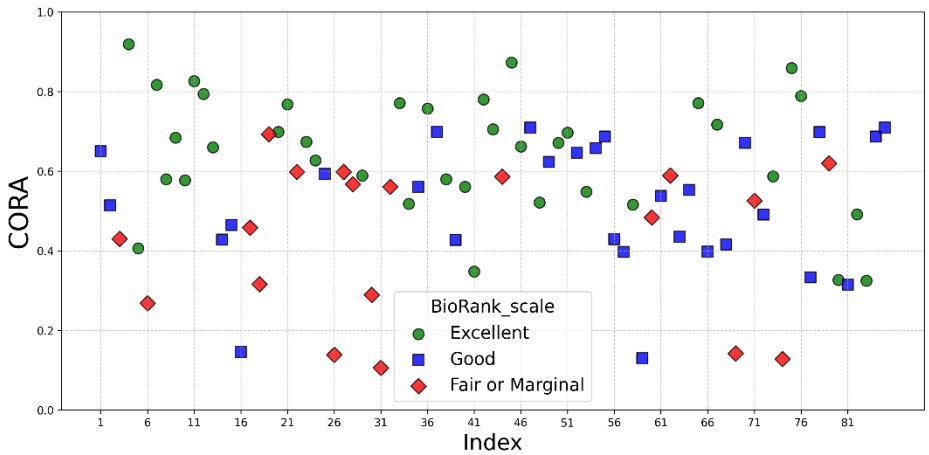


Fig. 11. CORA values coloured by BioRank scale.

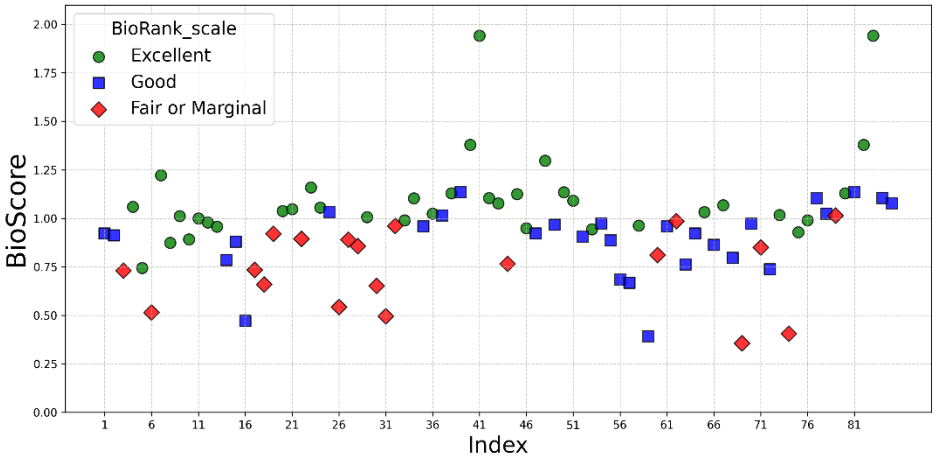


Fig. 12. BioScore values coloured by BioRank scale.

Table II shows the results of optimising the two thresholds based on the F1-score according to the three strategies described previously. In the brackets, the first threshold marks the transition from a *Fair or Marginal* grade to a *Good* grade, while the second threshold defines the boundary between a *Good* and an *Excellent* grade.

TABLE II
RESULTS OF 3 STRATEGIES TO DETERMINE CORA AND BIOSCORE GRADING SCALES

Strategy	CORA thresholds	BioScore thresholds	F1-Score BioRank-CORA	F1-Score BioRank-BioScore	F1-Score CORA-BioScore	F1-Score average (SD)
1	[0.32,0.67]	[0.66,0.98]	0.53	0.64	0.73	0.63 (0.10)
2	[0.29,0.52]	[0.66,0.82]	0.56	0.51	0.86	0.64 (0.19)
3	[0.29,0.52]	[0.66,0.98]	0.56	0.64	0.64	0.61 (0.05)

The average F1-scores for strategies 1, 2, and 3 are 0.63, 0.64, and 0.61, respectively. While strategy 2 achieves the highest average F1-score, it is the most unbalanced (F1-Score SD = 0.19), demonstrating a strong alignment between CORA and BioScore at the expense of the alignment between BioRank and the other two methods. Strategy 1 was ultimately selected as it provides a good compromise between balance and average F1-score. Additionally, thresholds are the highest for this strategy, which compensates for a general tendency observed in the data to grade curves overly positively.

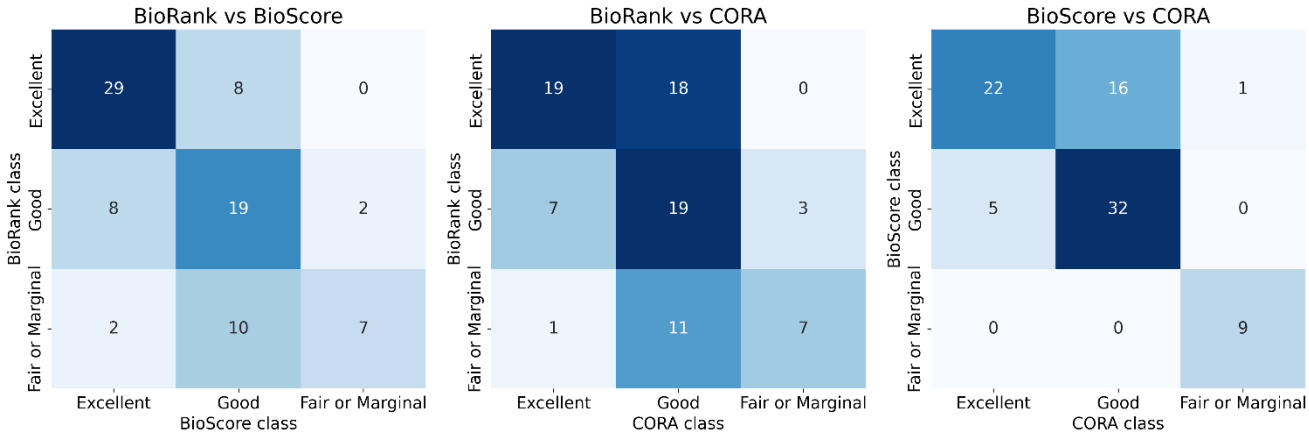


Fig. 13. Confusion matrix between the different metrics.

Figure 13 presents the confusion matrices corresponding to the thresholds derived from strategy 1. The figure clearly illustrates that, despite the optimisation, numerous cases are classified differently between BioRank and the other two methods. For instance, 10 signals rated as *Fair or Marginal* by BioRank were classified as *Good* by BioScore. However, only a few signals were classified two classes apart between the methods; notably, two signals labeled as *Excellent* by BioScore were graded as *Fair or Marginal* by BioRank.

In contrast, the class alignment between CORA and BioScore is the most consistent, with the primary discrepancy arising from 16 signals classified as *Good* by CORA and as *Excellent* by BioScore. BioScore generally tends to assign more favourable classifications compared to both other methods.

IV. DISCUSSION

BioScore differs fundamentally in that the dummy score is based on an average comparison with multiple individual PMHS curves. In contrast, CORA and BioRank aggregate the PMHS curves into a single average, allowing for a single curve comparison. Since the dummy-to-PMHS score is normalised by the PMHS-to-PMHS score, a scenario where there is a large difference between the dummy curve and the PMHS curves, but also high inter-PMHS variability, can be rated as favourably as a scenario where the dummy curve closely follows all PMHS curves but with low inter-PMHS variability. This effect is to some extent observed with BioRank, where the score is normalised by the SD of the PMHS curves. In the case of CORA, the corridor rating also accounts for dummy variability by incorporating 1SD corridors; however, the cross-correlation component of the rating does not account for it.

Reference [36] compared three biofidelity evaluation methods: CORA, BioRank, and the Sprague and Geers methods [37]. In that study, BioRank was implemented using the first version of the algorithm, which differs from the implementation in the current study, primarily because it used cumulative variance instead of absolute difference. Regression analyses revealed an R-squared value of 0.59 between the CORA corridor rating and BioRank, while the correlation between BioRank and the cross-correlation rating was nearly zero. Similarly, in the current study, the R-squared value between the CORA corridor rating and BioRank was 0.64, whereas the cross-correlation component of CORA again showed almost no correlation with BioRank. These findings confirm that BioRank aligns only with the corridor component of CORA and does not correlate with the cross-correlation component.

The results of this study highlighted that, aside from a relatively small number of outliers, CORA and BioScore appear to correlate well. Analysis of these outliers revealed that they all exhibited low PMHS-to-PMHS scores (<0.5), resulting in disproportionately high BioScore values. This effect is particularly pronounced when there are only a few PMHS curves, and the dummy curve closely matches one of them (see Figure 14 for an example). Interestingly, such cases also tend to receive positive ratings from BioRank, as the corresponding corridors are typically large.

In general, when PMHS curves exhibit poor correlation with one another, the validity of the results becomes questionable, regardless of the biofidelity evaluation method used. One way to address this issue could be to define a minimum PMHS-to-PMHS score below which biofidelity evaluation for a certain signal is not recommended. Another approach would be to introduce weighting factors based on the PMHS-to-PMHS score when calculating overall or regional biofidelity. This would ensure that cases with low PMHS-to-PMHS scores have less influence on the overall biofidelity assessment.

Conversely, when PMHS curves are tightly clustered, BioRank may assign a marginal grade to the dummy curve, even if it closely aligns with the PMHS curves, due to the method's reliance on narrow 1SD corridors in such cases. (see example Figure 15). CORA and BioScore, because they do not only rely on corridors but also consider shape and phase, tend to be more lenient with such signals.

As BioRank does not account for the shape, it also tends to grade signals which have a good shape but a bad amplitude more severely than CORA and BioScore (see example Figure 16).

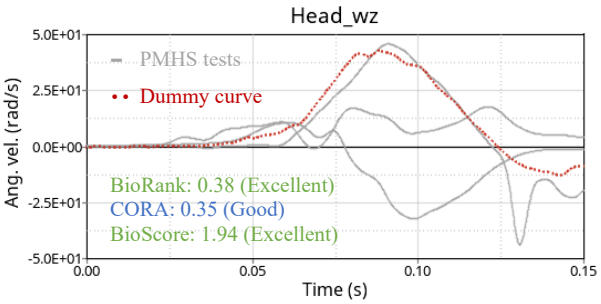


Figure 14. Example of disproportionately high BioScore value.

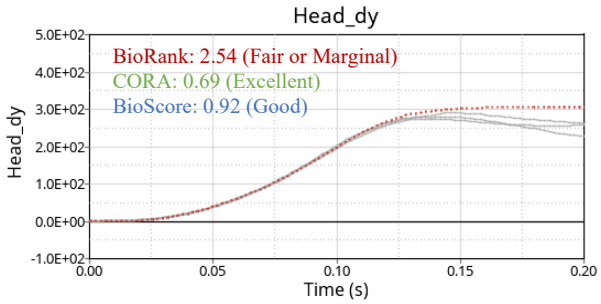


Figure 15. Example of a case where PMHS are very close to each other.

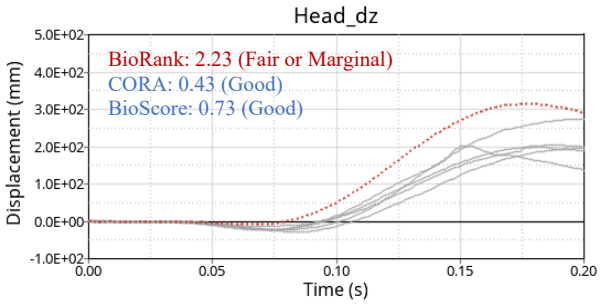


Figure 16. Example of a curve with good shape and bad amplitude.

To summarise, BioRank not accounting for shape leads to the overrating or underrating of certain signals. A common limitation of both CORA and BioRank is their reliance on averages and corridors. In this study, very few PMHS data are available for each signal, meaning that averages and corridors must be constructed from 3 to 5 PMHS curves, which can sometimes differ significantly. In such cases, the effect of averaging may be more detrimental for the CORA method than for BioRank, as CORA evaluates the difference in shape between the dummy curve and a potentially artificial PMHS average. This limitation is removed when using BioScore.

The CORA values calculated in this study are generally low. A previous study [38] suggests that for CORA values lower than 0.6, the sensitivity of the method is no longer sufficient to distinguish between signals. Although the implementation of CORA in this study differs, and individual curve-to-curve scores for BioScore are calculated according to the metric of ISO 18571 standard, it is still believed that a similar limitation may apply. This should be further investigated before recommending the use of CORA or BioScore for biofidelity evaluation, particularly because the lower threshold for CORA is below this value. It may for example be necessary to adjust the parameters of the CORA evaluation to improve the sensitivity of the algorithm at lower values. More broadly, any limitations of the ISO 18571 method will also apply to the CORA and BioScore ratings.

A general limitation of this study is the relatively small number of curves tested, as the analysis focused on sled tests, for which few PMHS curves are available. The presented results would need to be validated with additional cases containing more PMHS curves and regional or component tests.

Another limitation is that the grading scales for CORA and BioScore were established by aligning them with BioRank. Ideally, the grading scales for CORA and BioScore should be determined independently, possibly through expert judgement. Therefore, the thresholds proposed in this study should be further refined before being considered for future work. Alternatively, instead of grading scale, a probability or likelihood measure would be more useful, as it provides a quantifiable statistical metric and eliminates hard transition between classes. Additionally, the BioRank scale appears overly lenient; for example, a signal that deviates by 2 to 3 standard deviations from the mean is still classified as *Fair*. To avoid potential misinterpretation, the naming of the classification levels should likely be reconsidered.

It is worth noting that, at the time of writing this paper, Euro NCAP was in the process of publishing a technical bulletin on the use of HBMs in virtual testing. This bulletin introduces a new biofidelity evaluation method that utilises a corridor development approach called ARCGen [39]. One key advantage of this methodology is its ability

to evaluate hysteretic curves. In future works, this method should also be evaluated and compared to other approaches.

V. CONCLUSION

This study provides a comprehensive comparison of three biofidelity evaluation methods, BioRank, CORA, and BioScore, highlighting their respective strengths and limitations. The results demonstrate that the choice of method significantly impacts the biofidelity outcomes, with BioRank showing poor correlation with the other methods, while CORA and BioScore generally aligned well, aside from a few outliers. These outliers were primarily linked to low correlation between the individual PMHS curves, emphasising the importance of considering inherent variability within PMHS data when interpreting biofidelity scores.

BioRank and CORA's reliance on average curves and fixed corridors makes them sensitive to averaging issues, such as those caused by shape variations in PMHS curves, and corridor size effects, particularly when the corridors are very narrow. Conversely, BioScore's individual curve comparisons removes this limitation but may overestimate biofidelity when PMHS variability is high.

The development of grading scales for CORA and BioScore, aligned with BioRank, provides a practical framework for interpreting results, even if perfect alignment between methods remains elusive. The findings suggest that no single method is universally superior; but with some refinements and further validation, CORA and BioScore could improve on some of the limitations of BioRank.

Ultimately, this study underscores the need for further improvements and standardisation of biofidelity evaluation practices to ensure that human surrogates accurately represent real-world human responses, driving safer and more effective innovations in crash safety research.

VI. AKNOWLEDGEMENT

The Author would like to thank experts who have provided valuable feedback on the work and manuscript, namely Dr. Christan Gehre, Sebastian Fiedler, Dr. Giacomo Marini, Dr. Norbert Praxl, Dr. Debasis Sahoo, Philipp Wernicke and Michael Wolf.

VII. **REFERENCES**

[1] Rhule, H., Maltese, M., Donnelly, B., & Eppinger, R. H. (2002) Development of a new biofidelity ranking system for anthropomorphic test devices. Stapp Car Crash Journal, **46**: p.477-512.

[2] Gehre, C., Gades, H., & Wernicke, P. (2009) Objective rating of signals using test and simulation responses. *Proceedings of the 21st International Technical Conference on the Enhanced Safety of Vehicles (ESV),* 2009, Stuttgart, Germany.

[3] Barbat, S., Fu, Y., Zhan, Z., Yang, R.-J., & Gehre, C. (2013) Objective rating metric for dynamic systems. *Proceedings of the 23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2013, Seoul, South Korea.

[4] Parent, D., Craig, M., & Moorhouse, K. (2017) Biofidelity evaluation of the THOR and Hybrid III 50th percentile male frontal impact anthropomorphic test devices. Stapp Car Crash Journal, **61**: p.227–276.

[5] National Highway Traffic Safety Administration (NHTSA) *"Biomechanics database"*, https://www.nhtsa.gov/research-data/research-testing-databases#/biomechanics. [cited April 3, 2025]

[6] International Organization for Standardization. (2015) *ISO 6487:2015 - Measurement techniques in impact tests — Instrumentation*. Geneva, Switzerland.

[7] Rhule, H., Moorhouse, K., Donnelly, B., & Stricklin, J. (2009) Comparison of WorldSID and ES-2re biofidelity using an updated Biofidelity Ranking System. *Proceedings of the 21st International Technical Conference on the Enhanced Safety of Vehicles (ESV),* 2009, Stuttgart, Germany.

[8] Rhule, H., Donnelly, B., Moorhouse, K., & Kang, Y. S. (2013) A methodology for generating objective targets for quantitatively assessing the biofidelity of crash test dummies. *Proceedings of the 23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV),* 2013, Seoul, South Korea.

[9] Rhule, H. (2018) Improvements to NHTSA's Biofidelity Ranking System and application to the evaluation of the THOR 5th percentile dummy. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2018, Athens, Greece.

[10] Hagedorn, A., Stammen, J., Ramachandra, R., Rhule, H., Thomas, C., Suntay, B., Kang, Y. S., Kwon, H. J., Moorhouse, K., & Bolte, J. H. (2022) Biofidelity evaluation of THOR-50M in rear-facing seating configurations using an updated biofidelity ranking system. SAE International Journal of Transportation Safety, **10**(2): p.291–375.

[11] International Organization for Standardization. (2013) *ISO/TS 18571:2013 - Road vehicles — Objective rating metric for dynamic systems*. Geneva, Switzerland.

[12] Zhan, Z., Fu, Y., & Yang, R.-J. (2011) Enhanced Error Assessment of Response Time Histories (EEARTH) Metric and Calibration Process. *Proceedings of the SAE 2011 World Congress & Exhibition*, 2011, Detroit, USA.

[13] Park, G., Kim, T., Crandall, J. R., Arregui-Dalmases, C., & Luzón-Narro, J. (2013) Comparison of kinematics of GHBMC to PMHS on the side impact condition. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2013, Gothenburg, Sweden.

[14] Arun, M. W. J., Humm, J. R., Yoganandan, N., & Pintar, F. A. (2015) Biofidelity evaluation of a restrained whole body finite element model under frontal impact using kinematics data from PMHS sled tests. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2015, Lyon, France.

[15] Iwamoto, M., & Nakahira, Y. (2015) Development and validation of the Total HUman Model for Safety (THUMS) version 5 containing multiple 1D muscles for estimating occupant motions with muscle activation during side impacts. Stapp Car Crash Journal, **59**: p.53–90.

[16] Weaver, C., & Stitzel, J. D. (2015) Pelvic response of a Total Human Body Finite Element Model during simulated under-body blast impacts. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2015, Lyon, France.

[17] Arun, M. W. J., Umale, S., Humm, J. R., Yoganandan, N., & Pintar, F. A. (2016) Impact response evaluation of a restrained whole human body finite element model under far-side 90 and 60-degree impacts. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2016, Malaga, Spain.

[18] Feller, L., Kleinbach, C., Fehr, J., & Schmitt, S. (2016) Incorporating muscle activation dynamics into the Global Human Body Model. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2016, Malaga, Spain.

[19] Alshareef, A., Giudice, J. S., Forman, J., & Panzer, M. B. (2017) Evaluating the biofidelity of human brain finite element models using sonomicrometry data. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, Antwerp, Belgium.

[20] Östh, J., Mendoza-Vázquez, M., Lindner, A., Svensson, M. Y., & Brolin, K. (2017) The VIVA OpenHBM finite element 50th percentile female occupant model: Whole body model development and kinematics. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2017, Antwerp, Belgium.

[21] Kato, D., Nakahira, Y., Atsumi, N., & Iwamoto, M. (2018) Development of Human-Body model THUMS version 6 containing muscle controllers and application to injury analysis in frontal collision after brake deceleration. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2018, Athens, Greece.

[22] Song, E., Petit, P., & Uriot, J. (2018) Modelling of an adjustable generic simplified vehicle for pedestrian impact and simulations of corresponding reference PMHS tests using the GHBMC 50th percentile male pedestrian simplified model. Stapp Car Crash Journal, **62**: p.443–487.

[23] Iraeus, J., & Pipkorn, B. (2019) Development and validation of a generic finite element ribcage to be used for strain-based fracture prediction. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2019, Florence, Italy.

[24] Larsson, K. J., Pipkorn, B., Iraeus, J., Bolte IV, J. H., Agnew, A., Hu, J., Reed, M. P., & Sunnevang, C. (2019) Evaluation of the benefits of parametric human body model morphing for prediction of injury to elderly occupants in side impact. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2019, Florence, Italy.

[25] Perez-Repela, D., Markusic, C., Whitcomb, B., Pipkorn, B., Forman, J., & Crandall, J. (2019) Comparison of the simplified GHBMC to PMHS kinematics in far-side impact. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2019, Florence, Italy.

[26] Mroz, K., Östling, M., Richardson, R., Kerrigan, J., Forman, J., Gepner, B., Lubbe, N., & Pipkorn, B. (2020) Effect of seat and seat belt characteristics on the lumbar spine and pelvis loading of the SAFER human body model in reclined postures. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference,* 2020, Munich, Germany.

[27] Martynenko, O., Wochner, I., Nölle, L., Alfaro, E., Schmitt, S., Mayer, C., Mishra, A., Ghosh, P., Chitteti, R., Weber, J., González-Garcia, M., Beaugonin, M., & Vlachoutsis, S. (2021) Comparison of head-neck kinematics of different active human body models with experimental data. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2021, online.

[28] Pipkorn, B., Östh, J., Brynskog, E., Larsson, E., Rydqvist, L., Iraeus, J., Perez-Repela, D., & Jakobsson, L. (2021) Validation of the SAFER human body model kinematics in far-side impacts. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2021, online.

[29] Valdano, M., Jiménez-Octavio, R., Vives-Torres, C., López-Valdés, F., & Pipkorn, B. (2021) Assessment of MADYMO active human body model kinematics and dynamics by means of human volunteer responses at low-speed frontal impacts. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2021, online.

[30] Gepner, B., Perez-Repela, D., Forman, J., Östling, M., Pipkorn, B., & Kerrigan, J. (2022) Evaluation of GHBMC, THUMS, and SAFER human body models in frontal impacts in reclined postures. *Proceedings of the International Research Council on the Biomechanics of Injury (IRCOBI) Conference*, 2022, Porto, Portugal.

[31] Hsu, T., & Nusholtz, G. (2005). Considerations of biofidelity corridors for lateral impacts. *Proceedings of the SAE World Congress*, 2005, Detroit, USA.

[32] Nusholtz, G., Hsu, T., Shi, Y., Kochekseraii, S., & Luna, M. (2009) Creating representative curves from multiple time histories of vehicle, ATD, and biomechanics tests. *Proceedings of the Enhanced Safety of Vehicles (ESV) Conference*, 2009, Stuttgart, Germany.

[33] Nusholtz, G., Hsu, T., & Byers, L. (2007) A proposed side impact ATD biofidelity evaluation scheme using a cross-correlation approach. *Proceedings of the Enhanced Safety of Vehicles (ESV) Conference*, 2007, Lyon, France.

[34] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications*, R. Springer, New York.

[35] Powers, D. M. (2011) Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness & correlation. Information Processing & Management, **2**(1): p.37–63.

[36] Vavalle, N., Jelen, B., Moreno, D., Stitzel, J., & Gayzik, S. (2013) An evaluation of objective rating methods for full-body finite element model comparison to PMHS tests. Traffic Injury Prevention, **13**: p.87-94.

[37] Sprague, M. A., & Geers, T. L. (2004) A spectral-element method for modeling cavitation in transient fluid-

structure interaction. International Journal for Numerical Methods in Engineering, **60**(15): p.2467-2499.

[38] Kim, S. A. (2011) TU Berlin, *Bewertung der Korrelation von Messsignalen - Sensitivitätsanalyse der Steuerparameter der Bewertungssoftware CORA*. Berlin, Germany.

[39] Hartlen, D. C., & Cronin, D. S. (2022) Arc-Length Re-Parametrization and Signal Registration to Determine a Characteristic Average and Statistical Response Corridors of Biomechanical Data. Frontiers in Bioengineering and Biotechnology, **10**.

APPENDIX 1: Details on curve rating methods

BioRank: Calculation of CCSD and DCAD

The CCSD is calculated by summing up the standard deviation values from the cadaver mean curve over the time range of the evaluation as illustrated in Figure 17. The DCAD is calculated by summing up the absolute difference between the PMHS average and the dummy curve (Figure 18).

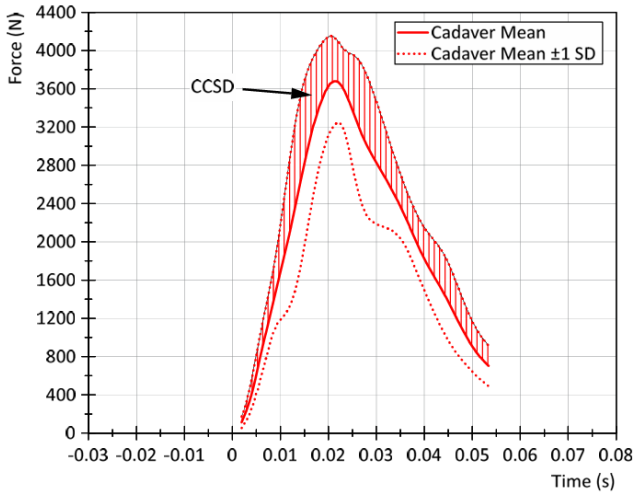


Figure 17: Illustration of the calculation of CCSD [9].

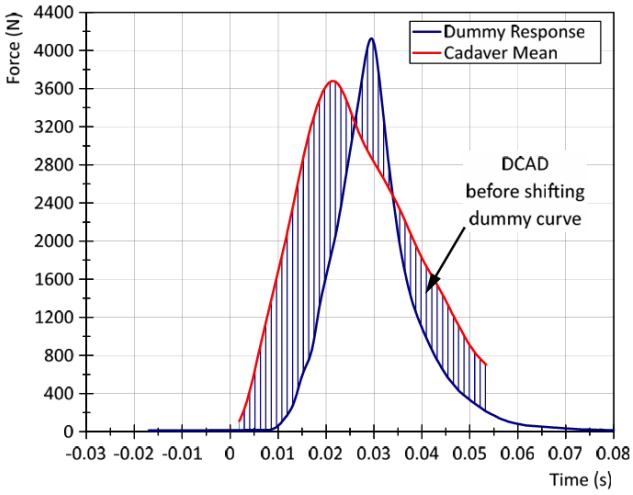


Figure 18: Illustration of the calculation of DCAD [9].

EEARTH method

The rating consists of three components: a phase error, a magnitude error, and a slope error. A particular feature of the method is the use of a DTW algorithm allowing the evaluation of each error component while minimizing the effect of the others. A cross-correlation function is used to find the optimal number of timestep shift, n between the two signals. The error is then calculated using the following equation:

$$Error_{phase} = e^{((n-c)/r)} \tag{8}$$

c and r are parameters which define the acceptable phase shift (shift below which the error is considered negligeable) and the rate at which the phase error rises if it rises above the acceptable phase shift. The exponential formulation is aimed at discriminating between negligeable and significant phase shifts. Values of 15 and 20 are advised for crash application for c and r respectively.

Before the magnitude error is computed, the DTW algorithm is applied to minimize the effect of phase and frequency errors. The cost function for the DTW integrates magnitude and slope errors:

$$d[i,j] = \left(\left(a_i^{ts} - b_i^{ts}\right)^2 + \left(t_i - t_j\right)^2\right)\left|\left(\frac{dA^{ts}}{dt}\right)_{t=ti} - \left(\frac{dB^{ts}}{dt}\right)_{t=tj}\right| \tag{9}$$

where ts is used to denote time-shifted time histories.

Once the cost function is minimized between the two signals, the magnitude error is calculated using the equation below:

$$Error_{magnitude} = \frac{\sum|a_i - b_i|}{\sum b_i^2} \tag{10}$$

The slope error is computed on the derivate of the time histories. The derivate is applied to the phase shifted signals and the error is calculated as follows:

$$Error_{slope} = \frac{\sum|\partial a_i/\partial t - \partial b_i/\partial t|}{\sum(\partial b_i/\partial t)^2} \tag{11}$$

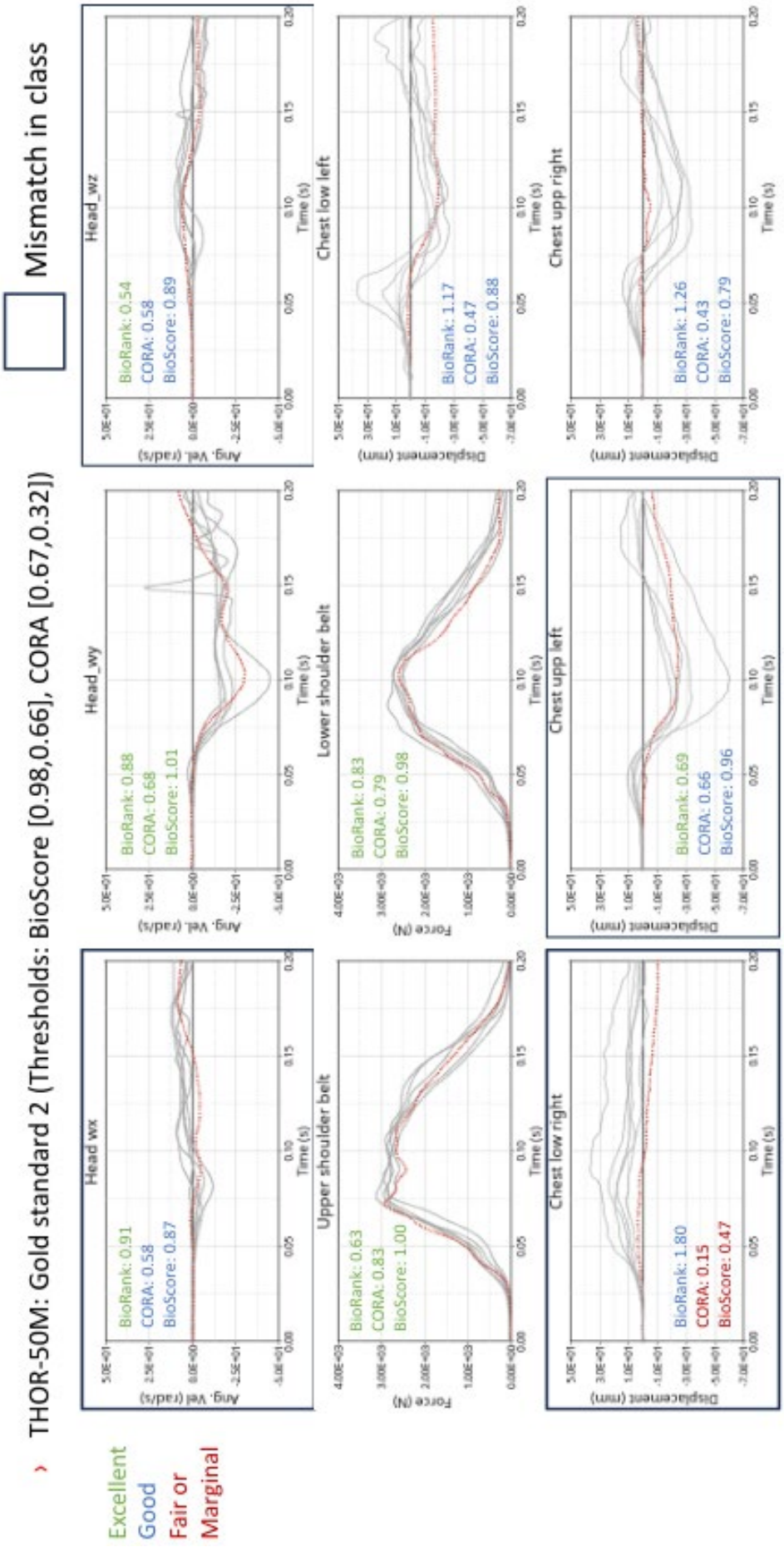
The three errors are combined in a single score using a weighted sum. The weights applied to the different errors were obtained using experts' opinions. The same principles are used for both EARTH and EEARTH, but in the enhanced version, three points were improved:

- The method used to obtain the weight for each error uses an integrated calibration process and the ratings are between 0 and 1
- The cost function for the DTW uses only the distance instead of both slope and distance
- The slope rating is performed on the time shifted derivates without application of the DTW

The last two points are aimed at improving the robustness of the algorithm to differences in sampling rates.

APPENDIX 2

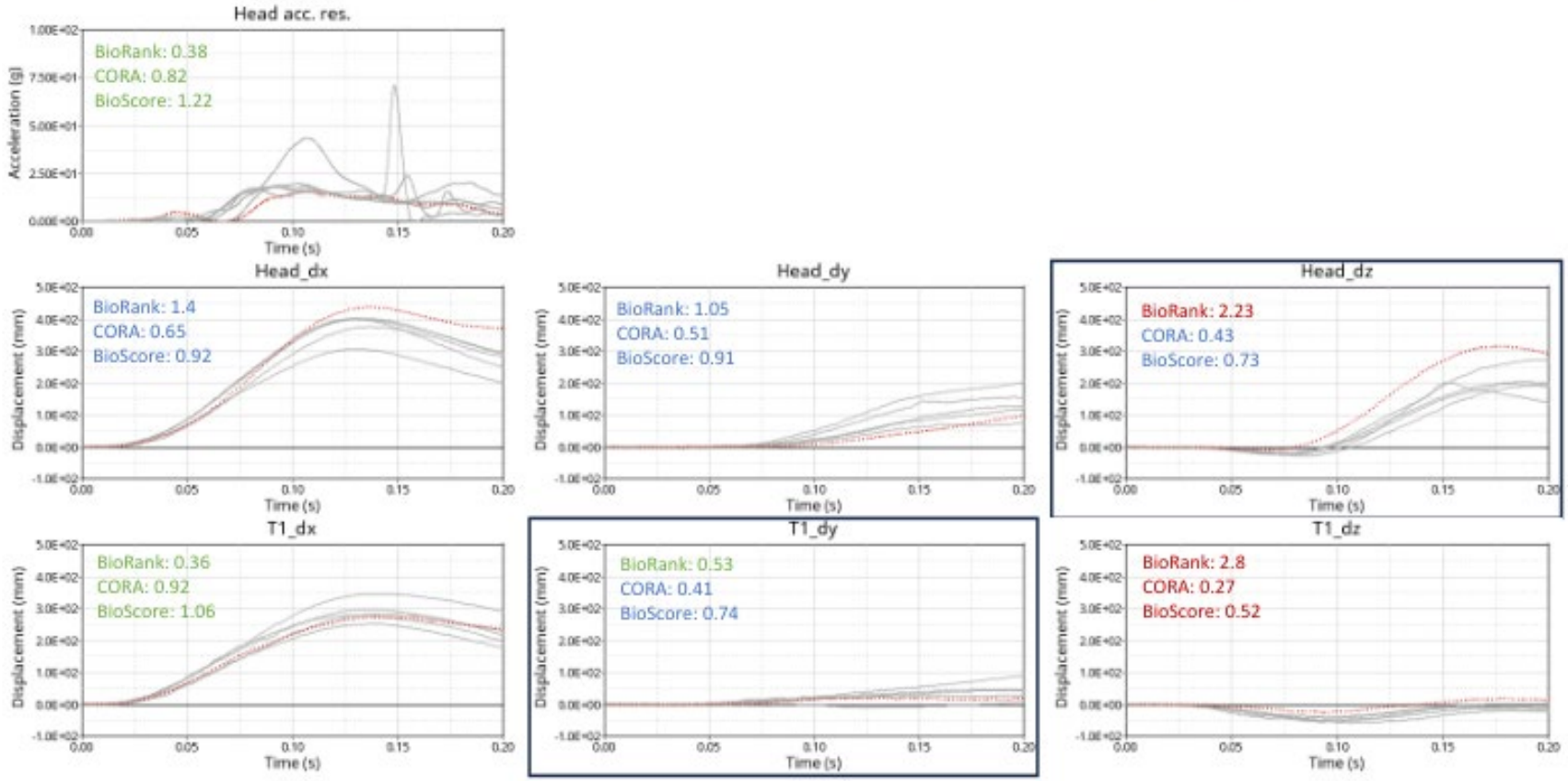
All curves and corresponding ratings

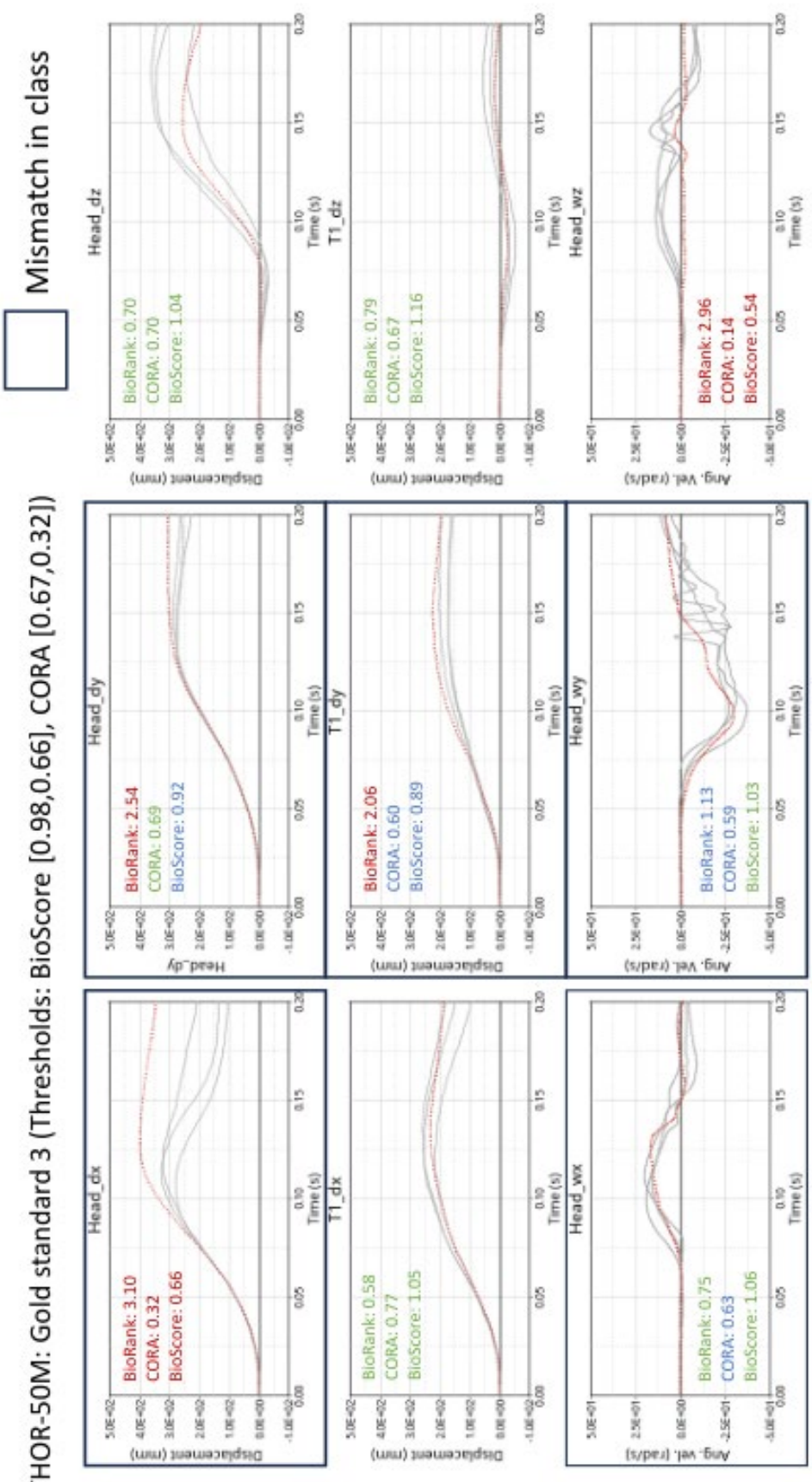


› THOR-50M: Gold standard 2 (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])

Mismatch in class

Excellent
Good
Fair or
Marginal

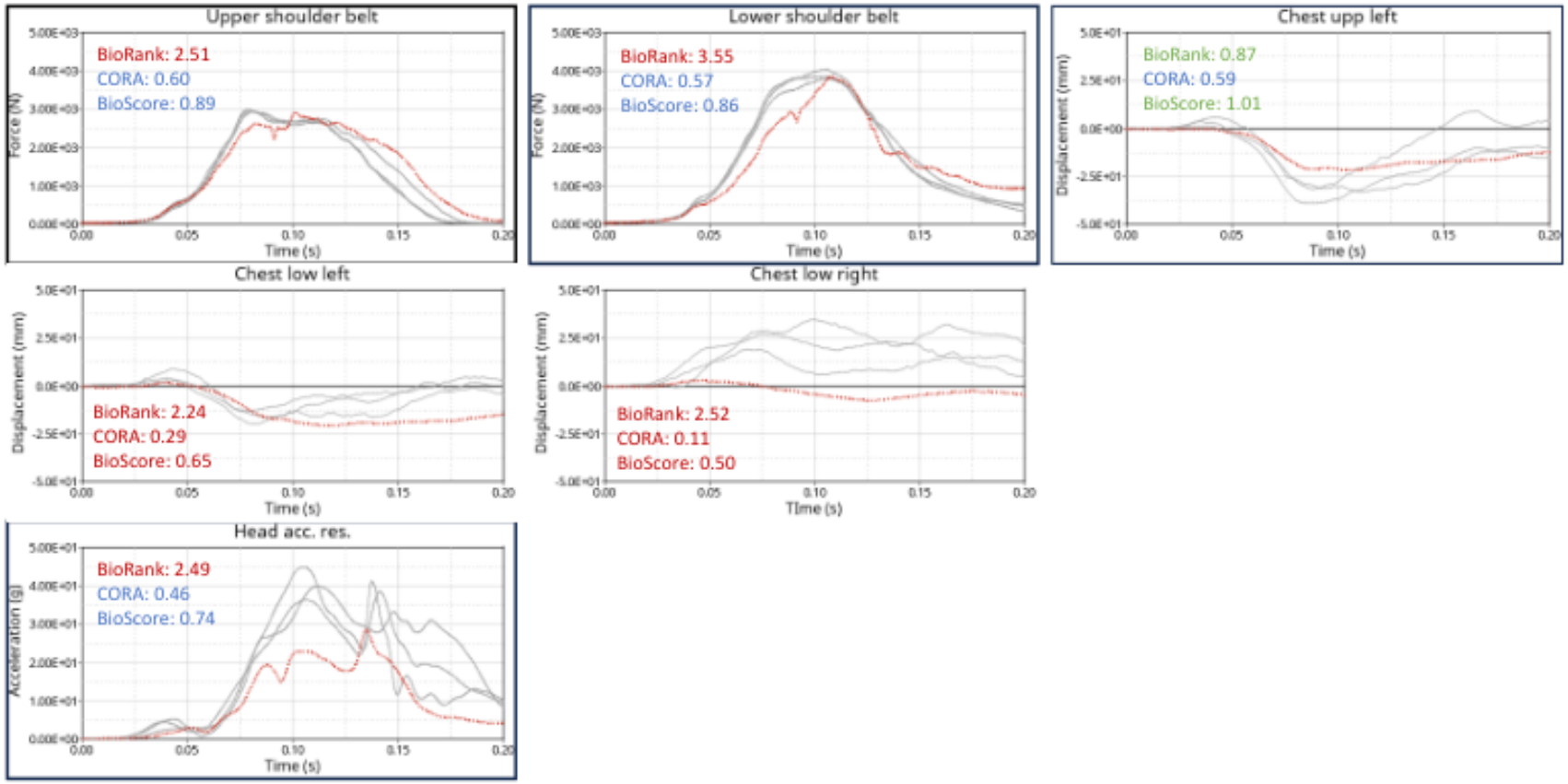



THOR-50M: Gold standard 3 (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])
Mismatch in class
Excellent
Good
Fair or
Marginal
Head_dx
BioRank: 3.10
CORA: 0.32
BioScore: 0.66
T1_dx
BioRank: 0.58
CORA: 0.77
BioScore: 1.05
Head_dy
BioRank: 2.54
CORA: 0.69
BioScore: 0.92
T1_dy
BioRank: 2.06
CORA: 0.60
BioScore: 0.89
Head_dz
BioRank: 0.70
CORA: 0.70
BioScore: 1.04
T1_dz
BioRank: 0.79
CORA: 0.67
BioScore: 1.16
Head_wx
BioRank: 0.75
CORA: 0.63
BioScore: 1.06
Head_wy
BioRank: 1.13
CORA: 0.59
BioScore: 1.03
Head_wz
BioRank: 2.96
CORA: 0.14
BioScore: 0.54
Displacement (mm)
Ang. Vel. (rad/s)
Time (s)

› THOR-50M: Gold standard 3 (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])

Mismatch in class

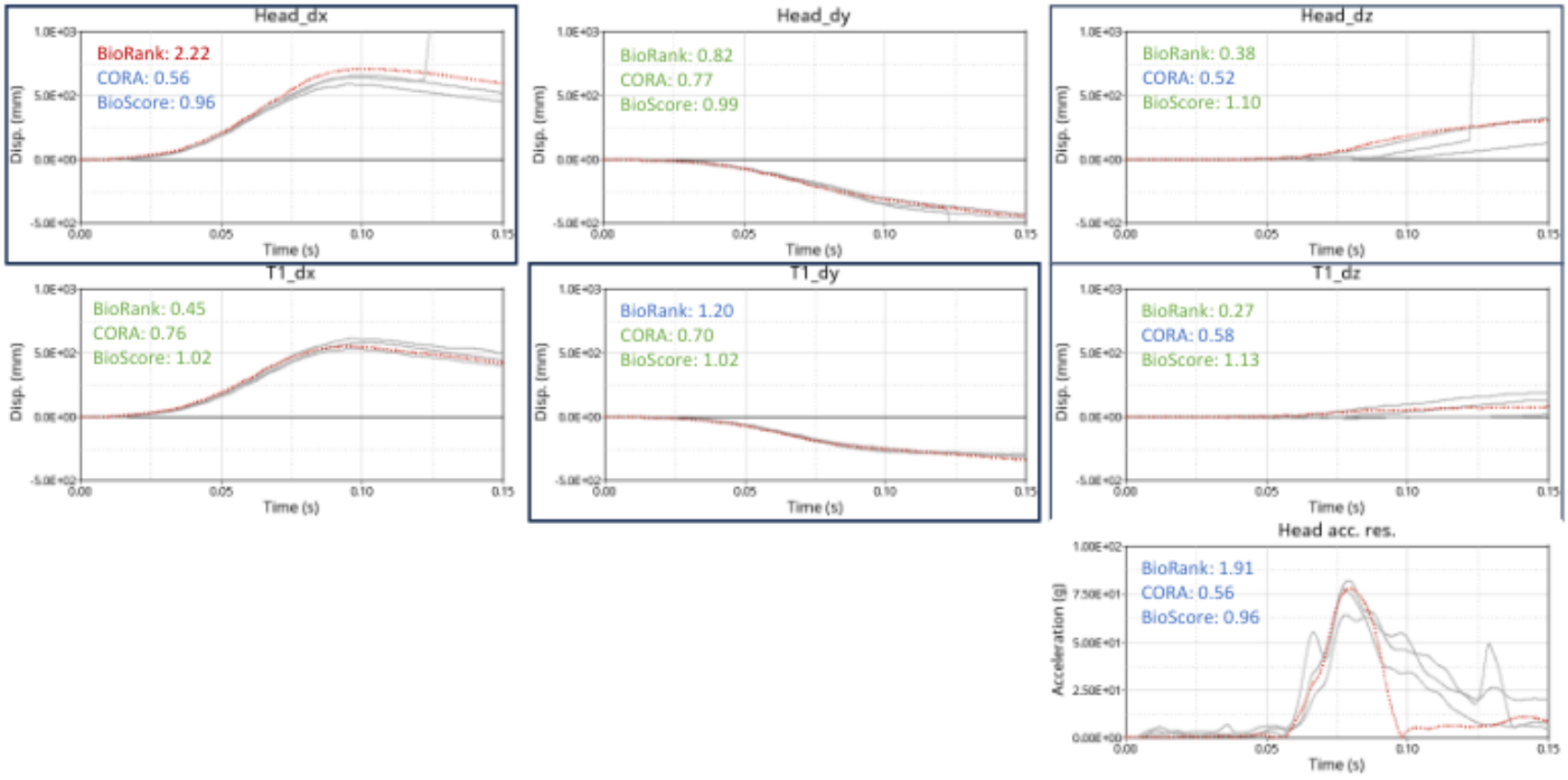
Excellent
Good
Fair or
Marginal



› THOR-50M: Far-side oblique (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])

Mismatch in class

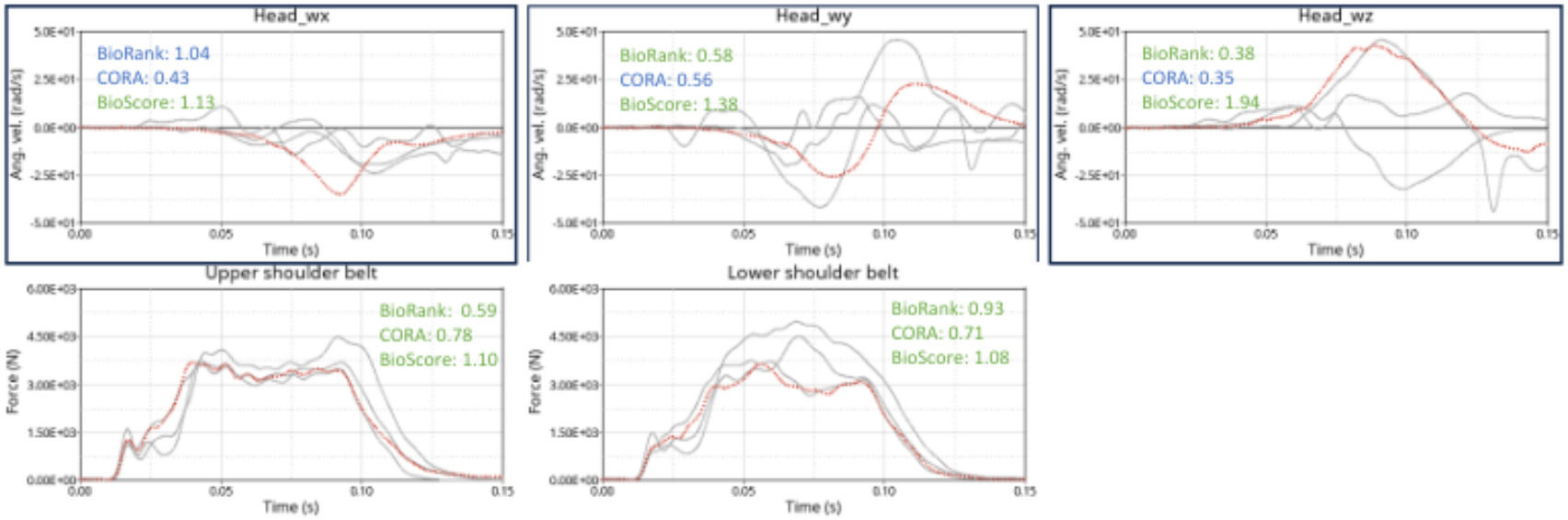
Excellent
Good
Fair or
Marginal

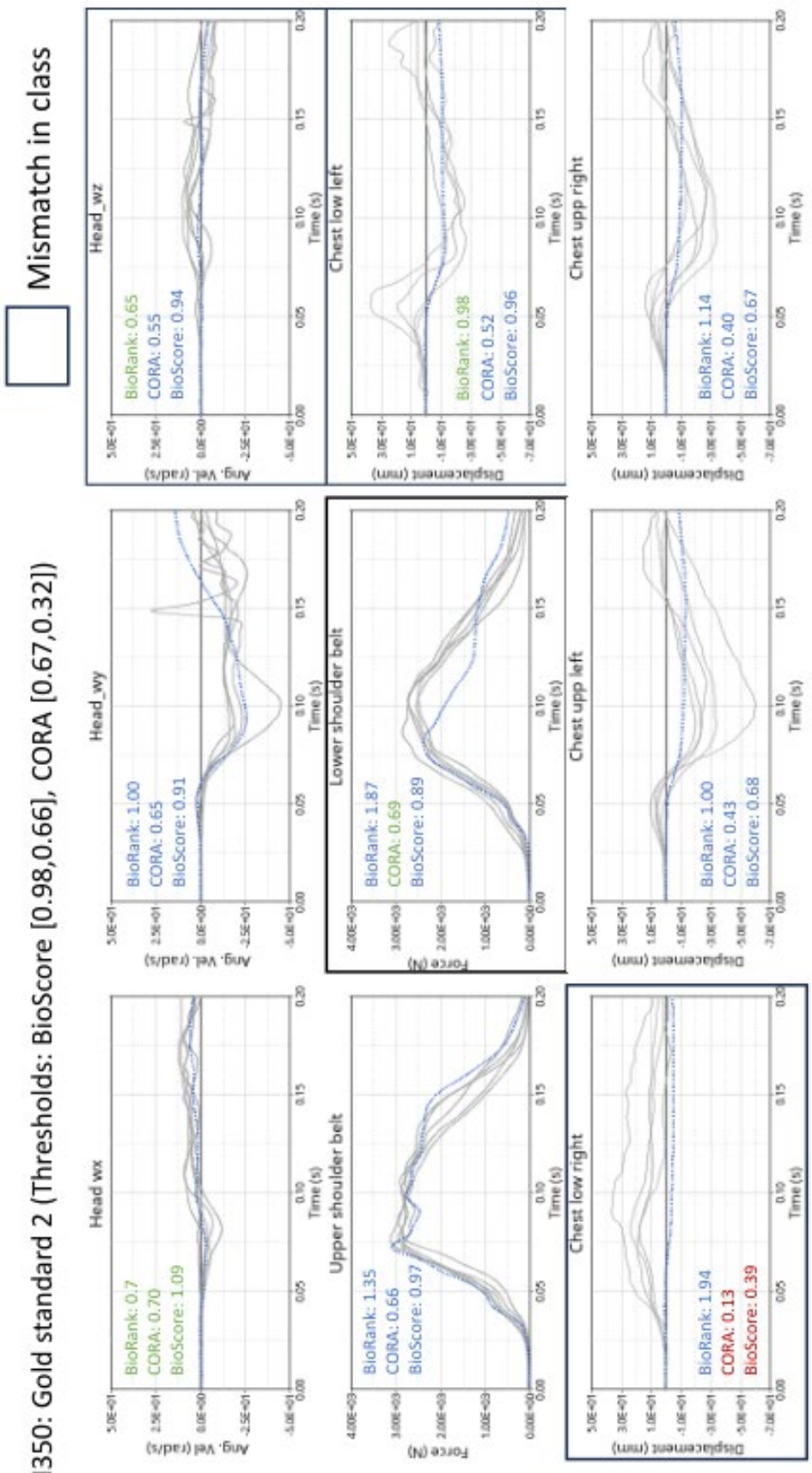


› THOR-50M: Far-side oblique (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])

Mismatch in class

Excellent
Good
Fair or
Marginal

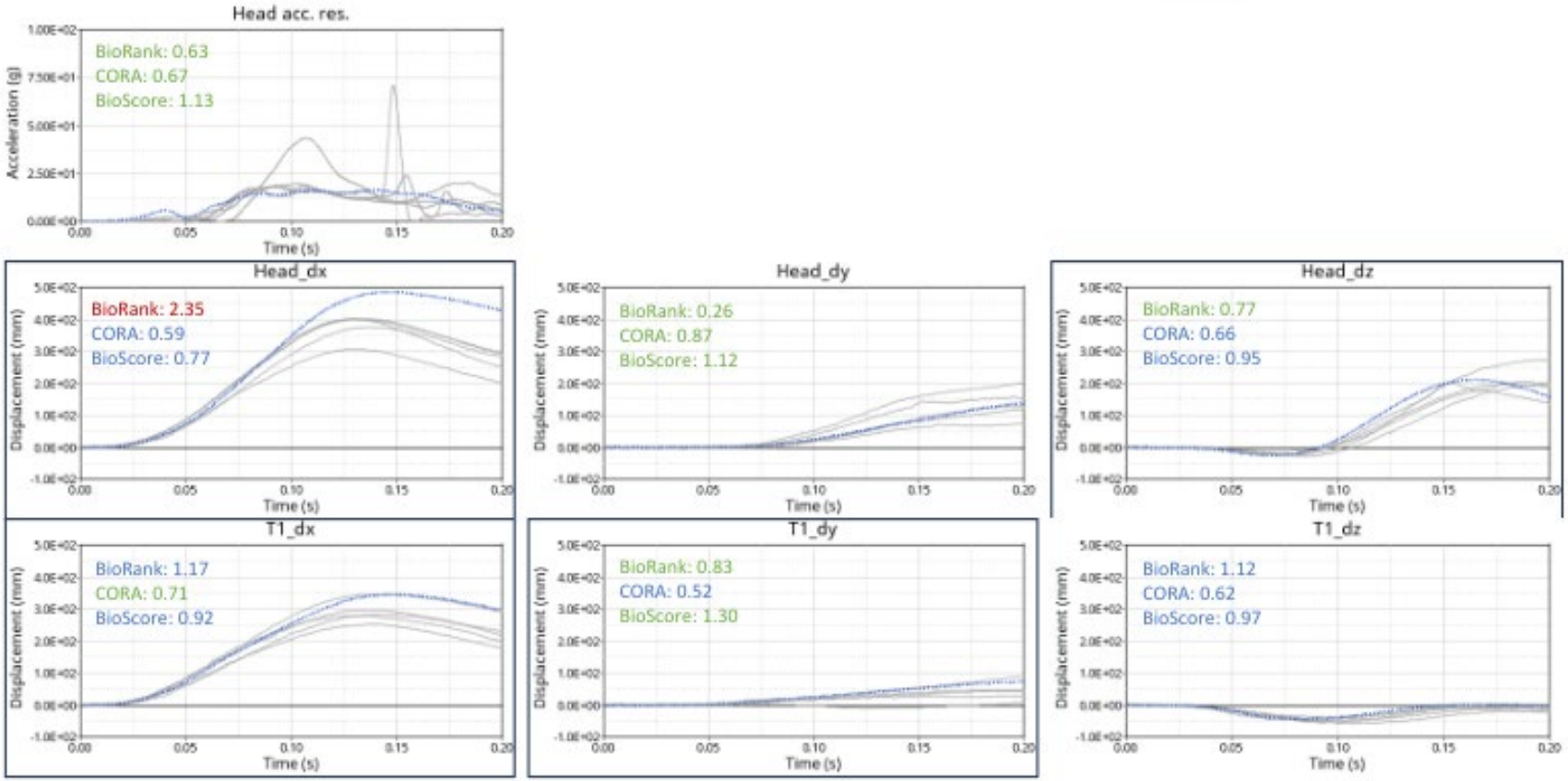


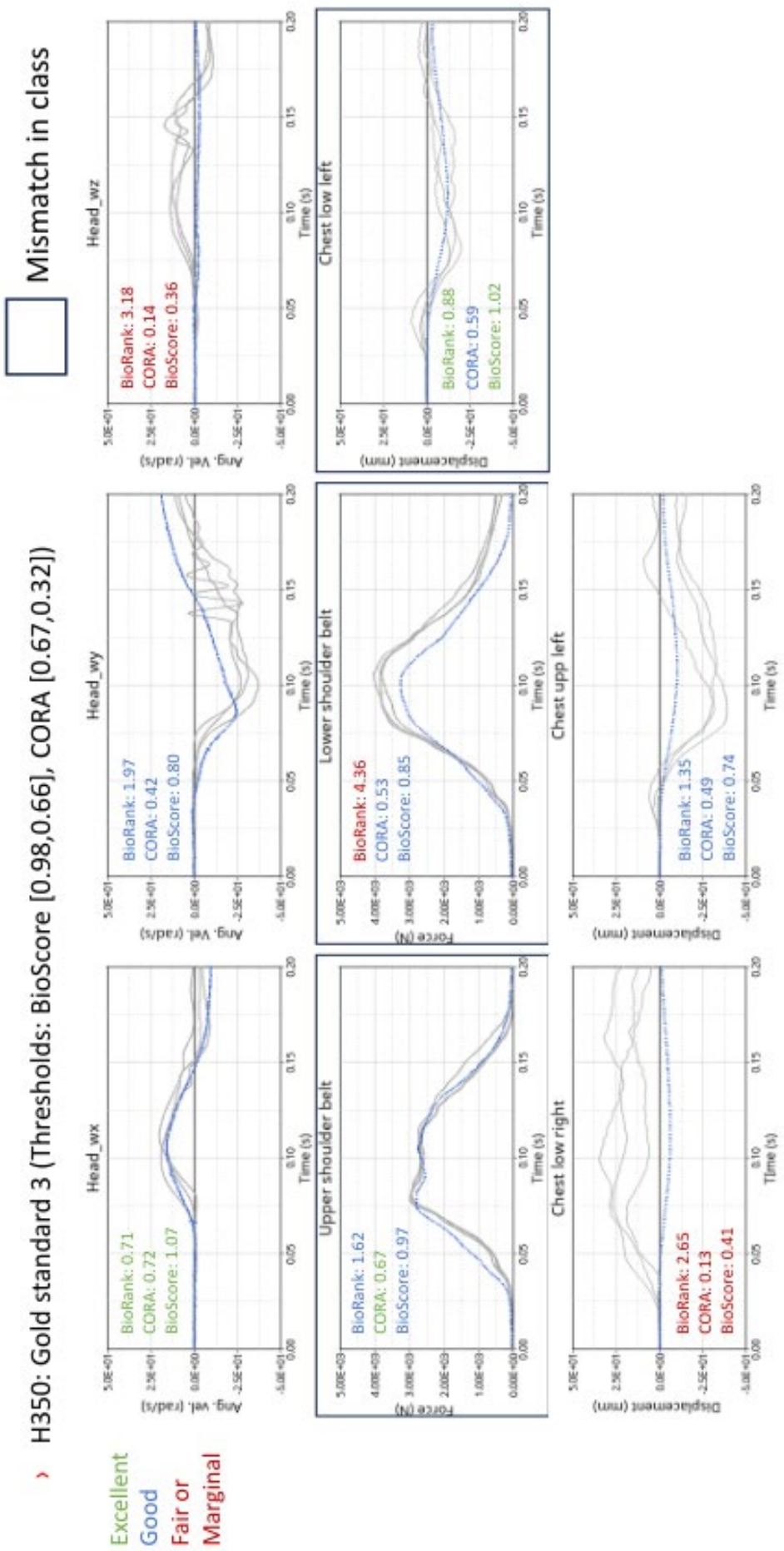

H350: Gold standard 2 (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])
Mismatch in class
Excellent
Good
Fair or
Marginal
Head_wz
BioRank: 0.65
CORA: 0.55
BioScore: 0.94
Chest low left
BioRank: 0.98
CORA: 0.52
BioScore: 0.96
Chest upp right
BioRank: 1.14
CORA: 0.40
BioScore: 0.67
Head_wy
BioRank: 1.00
CORA: 0.65
BioScore: 0.91
Lower shoulder belt
BioRank: 1.87
CORA: 0.69
BioScore: 0.89
Chest upp left
BioRank: 1.00
CORA: 0.43
BioScore: 0.68
Head wx
BioRank: 0.7
CORA: 0.70
BioScore: 1.09
Upper shoulder belt
BioRank: 1.35
CORA: 0.66
BioScore: 0.97
Chest low right
BioRank: 1.94
CORA: 0.13
BioScore: 0.39
Ang. Vel. (rad/s)
Force (N)
Displacement (mm)
Time (s)

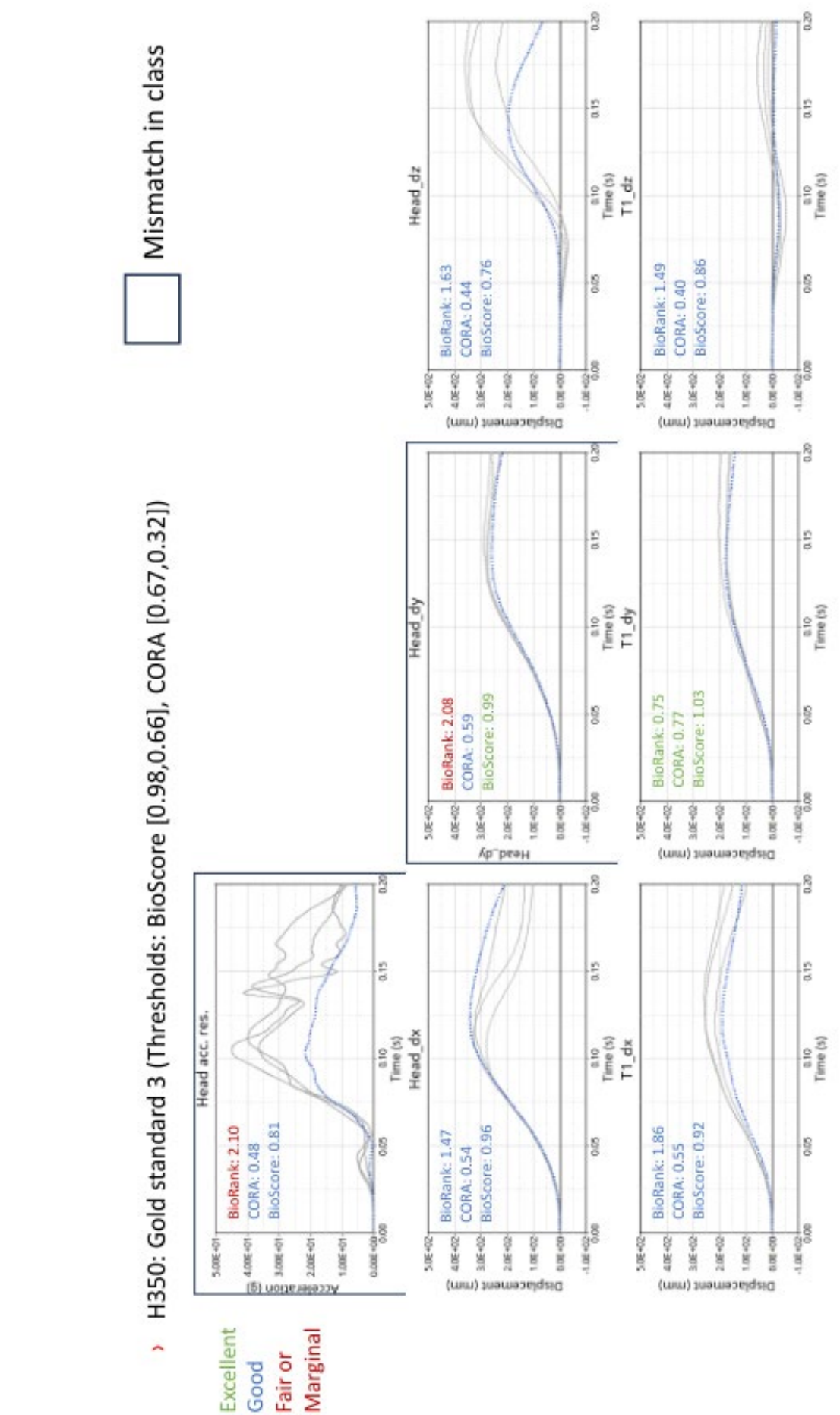
› H350: Gold standard 2 (Thresholds: BioScore [0.898,0.66], CORA [0.67,0.32])

Mismatch in class

Excellent
Good
Fair or
Marginal



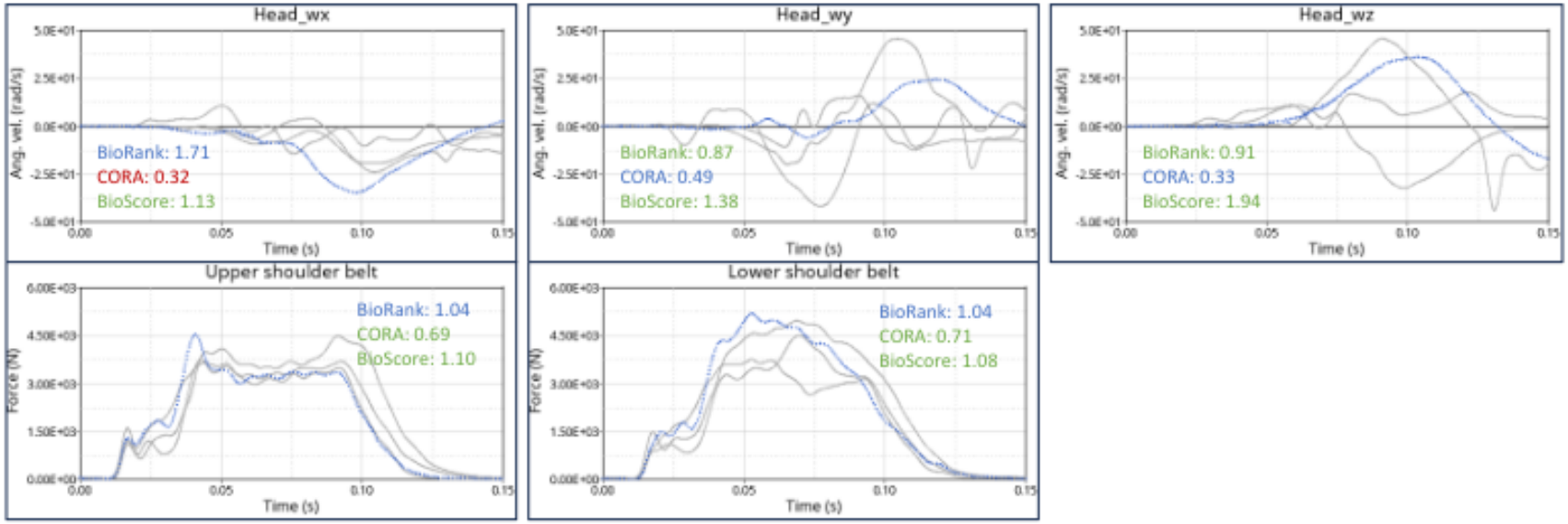

H350: Gold standard 3 (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])
Mismatch in class
Excellent
Good
Fair or
Marginal
Head_wx
BioRank: 0.71
CORA: 0.72
BioScore: 1.07
Head_wy
BioRank: 1.97
CORA: 0.42
BioScore: 0.80
Head_wz
BioRank: 3.18
CORA: 0.14
BioScore: 0.36
Upper shoulder belt
BioRank: 1.62
CORA: 0.67
BioScore: 0.97
Lower shoulder belt
BioRank: 4.36
CORA: 0.53
BioScore: 0.85
Chest low left
BioRank: 0.88
CORA: 0.59
BioScore: 1.02
Chest low right
BioRank: 2.65
CORA: 0.13
BioScore: 0.41
Chest upp left
BioRank: 1.35
CORA: 0.49
BioScore: 0.74
Ang. Vel. (rad/s)
Force (N)
Displacement (mm)
Time (s)


H350: Gold standard 3 (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])
Mismatch in class
Excellent
Good
Fair or
Marginal
Head acc. res.
BioRank: 2.10
CORA: 0.48
BioScore: 0.81
Head_dx
BioRank: 1.47
CORA: 0.54
BioScore: 0.96
Head_dy
BioRank: 2.08
CORA: 0.59
BioScore: 0.99
Head_dz
BioRank: 1.63
CORA: 0.44
BioScore: 0.76
T1_dx
BioRank: 1.86
CORA: 0.55
BioScore: 0.92
T1_dy
BioRank: 0.75
CORA: 0.77
BioScore: 1.03
T1_dz
BioRank: 1.49
CORA: 0.40
BioScore: 0.86
Time (s)
Displacement (mm)
Acceleration (g)

› H350: Far-side oblique (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])

Mismatch in class

Excellent
Good
Fair or
Marginal



› H350: Far-side oblique (Thresholds: BioScore [0.98,0.66], CORA [0.67,0.32])

Mismatch in class

Excellent
Good
Fair or
Marginal