COMPARISON OF OBJECTIVE RATING TECHNIQUES VS. EXPERT OPINION IN THE VALIDATION OF HUMAN BODY SURROGATES

Matthew L. Davis
Wake Forest School of Medicine
Virginia Tech-Wake Forest University Center for Injury Biomechanics
United States

Bharath Koya
Wake Forest School of Medicine
Virginia Tech-Wake Forest University Center for Injury Biomechanics
United States

Jeremy M. Schap
Wake Forest School of Medicine
Virginia Tech-Wake Forest University Center for Injury Biomechanics
United States

Fang-Chi Hsu
Wake Forest School of Medicine
Department of Biostatistical Sciences
United States

F. Scott Gayzik
Wake Forest School of Medicine
Virginia Tech-Wake Forest University Center for Injury Biomechanics
United States

Paper Number 17-0141

ABSTRACT

Objective evaluation (OE) methods provide quantitative insight into how well human body models (HBMs) predict a biomechanical response. Two techniques for this purpose are CORA and the ISO/TS 18571 standard. These ostensibly objective techniques have differences in their algorithms that may lead to discrepancies when interpreting model performance. The objectives of this study were 1) to apply both techniques to a biomechanical dataset from a HBM, and compare the scores and 2) conduct a survey of subject matter experts (SMEs) to determine which OE method compares more consistently with SME interpretation. The GHBMC average male HBM was used in five simulations of biomechanics experiments, producing 58 time history curves. Because both techniques produce phase, magnitude, and shape scores, 174 pairwise comparisons were made. ISO had lower average scores for each component rating metric than CORA, indicating a stricter evaluation. Correlations between CORA and ISO were strongest for phase (R^2=0.66) and weakest for shape (R^2=0.27). Statistical analysis revealed significant differences between the two OE methods for each component rating metric. SMEs (n=40) were then surveyed to provide intuitive scoring of how well the computational traces matched the experiments. SME interpretation was found to statistically agree with the ISO shape and phase metrics, but was significantly different than the ISO magnitude rating. SME interpretation agreed with the CORA magnitude rating. The finding of the study suggests a mixed approach to reporting objective ratings, using the magnitude method in CORA and the ISO shape and phase methods.

INTRODUCTION

The use of computational modeling has become an important aspect of the development process in the automotive and defense industries. Prior to production, products are often tested using a variety of computer programs to evaluate their performance. These simulations evaluate aspects of the design process ranging from structural crash-worthiness [1] to occupant protection and injury risk mitigation [2]. A growing component of these types of analyses includes the use of computational human body surrogates. These simulations can include a variety of models, such as rigid body models [3], anthropomorphic test devices (ATDs) [4], or full human body models (HBMs) [2,5]. Due to the reduced cost of running these simulations, as well as the large amount data that can extracted from them, these types of simulations offer a valuable supplement to physical testing. However, in order for these models to yield meaningful data, they must be carefully validated. How closely a model matches an experiment is a key piece of information for modelers. For example, HBMs are commonly compared against mean response and corridor biomechanical data obtained from Post-Mortem Human Subject (PMHS) testing [6]. For the sake of validation, a quantitative comparison that leads to an unambiguous interpretation of the model performance, taking into account the biological variation of specimens, is paramount when characterizing the biofidelity of a model. These objective comparisons offer a robust means of evaluating the performance of a model throughout the course of development.

Objective Evaluation (OE) methods seek to replace the subjectivity inherent in the validation process with a numerical score that provides quantitative insight into how well a human surrogate predicts a biomechanical response. While there are many techniques for this purpose [7-9], two commonly applied methods are Gehre et al.'s CORA software [10] and the ISO/TS 18571 standard [11]. The advantage of these techniques is that they evaluate individual components of the curve to provide a more complete comparison of time-history signals. While both techniques evaluate similar aspects of the signals, there are several differences between the inherent algorithms of the methods that can lead to different interpretations of the results. It is important to understand how these differences can be interpreted and the effect they can have on model validation. While these techniques are broadly used [2,12,13], they have not been directly compared.

As such, the objectives of this study are two-fold: 1) compare the results of CORA and ISO/TS 18571 OE techniques applied to a set of biomechanical data derived from a human body finite element model, and 2) conduct a survey of subject matter experts (SMEs) to determine which of these OE methods, if either, compares more consistently with SME interpretation. The goal of this work is to evaluate how results from these techniques can influence the interpretation of model validity, and if these interpretations agree with real world expert interpretation.

METHODS

The Global Human Body Models Consortium (GHBMC) average male occupant (M50-O v4.4) finite element model was selected for use in this study. The model was developed based on a multi-modality medical image and external anthropometry dataset of a volunteer representing a 50th percentile male in terms of height (174.9 cm) and weight (78.6 ± 0.77 kg). The development and application of this dataset was described by Gayzik et al. [14]. Once developed, the model underwent validation simulations at both the regional [15-18] and full body levels [19-21]. More information on the development of the model can be found in the GHBMC M50-O user's manual [22].

Simulations

To obtain outputs representing a range of impact conditions and directions, the model was run through five simulations representing physical biomechanics experiments. These simulations consisted of both localized, rigid hub impacts and full body sled cases. The rigid hub simulations included an oblique thoracoabdominal impact [23], a frontal abdominal impact [24], and a lateral pelvis impact [25]. The full body sled cases represented a lateral impact into fixed steel plates [26,27] and a frontal sled test configuration [28]. All simulations were run using LS-Dyna v6.1.1, rev. 78769 on a Linux Red Hat 6 high performance computing system (the Distributed Environment for Academic Computing, or DEAC cluster) maintained at Wake Forest University.

The thoracoabdominal impact employed a 23.4 kg cylindrical hub impactor with a 15 cm diameter and a nominal impact velocity of 6.7 m/s [23]. The impact location was 7.5 cm below the xipohoid process at 60° from anterior. Model data were compared to the mean experimental force vs. time signal in order to evaluate the OE techniques.

The abdominal impact consisted of a 2.5 cm diameter, 48 kg bar impacting at 6.0 m/s. This was a free-back impact occurring at the level of the umbilicus (approximately L3) [24]. The force of the impact was measured as the contact force of the rigid bar. Model data were compared to the mean experimental force vs. time curve.

The pelvic impact simulated a square-faced impactor weighing 16 kg impacting with 800 J of energy. This required giving the impactor a 10 m/s velocity normal to the sagittal plane. The pelvis impactor contacted the trochanter and iliac crest at 90°, according to the literature [25]. Similar to the other rigid impact simulations, the contact force of the impacting plate was used to obtain force data. This contact force was compared to the mean experimental force vs. time curve to facilitate OE technique comparison.

The lateral sled test was modeled as a 6.7 m/s impact using a Heidelberg-type sled [26,27]. The impact environment included a flat rigid wall as a backrest, a Teflon seat, and five rigid impacting plates located at the shoulder, thorax, abdomen, pelvis, and knee. Torso forces were obtained as the sum of the shoulder, thorax, and abdomen forces. The pelvis force was measured as the contacting force at the pelvis plate. For both the torso and pelvis outputs, the model responses were compared to the mean experimental force vs. time data.

The frontal sled case was modeled as per Shaw et al. [28]. This simulation represented a frontal impact with on overall change in velocity of 40 kph. The simplified buck used in the simulation was modeled as a rigid body. Belt properties were developed to match experimental conditions (26 kN of force at 7% strain) and no pretensioners or load-limiters were included. A foam knee bolster was also included to restrict motion of the lower extremities in the model. Prior to simulation, the model was gravity settled for 100 ms to obtain realistic flesh contours within the buck. With regards to outputs, both kinetic and kinematic responses were obtained for comparison to experimental values. With the exception of chest deflection data, all kinematics were reported in the global coordinate system. Reaction forces at the knee bolster and foot rests were also recorded in the global coordinate system and then transformed into a local coordinate system per the literature [29]. Resultant belt force data were obtained to represent the responses of the upper and lower shoulder belt and the outer lap belt. All data extracted from the model were compared against the average of experimental PMHS tests [28].

Objective Evaluation

While the model validity and accurate representation of the described biomechanical simulations is paramount, the goal of this study is to see how, when presented with identical comparison cases, the CORA and ISO techniques interpret model performance. To facilitate this comparison, all model data were output in binary files from LS-Dyna and were recorded at a sampling rate of 10 kHz. Post-processing of the data was performed in OASYS T-His (Ove Arup SYStems, Solihull, UK) and Matlab R2013 (MathWorks, Natick, MA). Force data were filtered using an SAE CFC 600 filter and kinematic data were not filtered.

In order to effectively source discrepancies between the two OE techniques, it is important to understand how each component of the rating metric is calculated. Detailed descriptions of the algorithms for each technique can be found in the literature [30,31]. However, as a foundation for comparison, each component of the CORA and ISO techniques is briefly described.

CORA

The CORA rating metric is a set of algorithms comprised of two independent sub-rating schemes: a corridor score and a cross-correlation score [10]. A complete description of this technique can be found in the literature [30]. The software was developed to calculate the level of correlation between two non-ambiguous signals and return a total score ranging from 0 to 1, where a 1 would indicate good correlation and a 0 would be a poor match based on defined tolerances. The default settings as recommended by the software provider were used in this analysis, with the exception of the phase interval, which is described below.

The corridor rating is designed to evaluate the deviation between the signals using a set of fixed-width or user-defined (i.e. experimentally reported) inner and outer corridors. If the model curve is within the inner corridor, the resulting score is a 1. If the model curve falls between the inner and outer corridor, the result is between 0 and 1 based on an interpolation score. If the signal is outside of the outer corridor, the result is a 0. While this technique gives a valuable global picture of model performance, a disadvantage of this approach is that phase differences between the model signal and the experimental data can lead to poor scores.

The cross correlation method analyzes three aspects of the signal in order to reduce the relative disadvantages of using only the corridor score: phase, shape, and magnitude. First, the algorithm attempts to eliminate differences in phasing by shifting the model curve by multiples of Δt. Then, for each shifted state, the program calculates a cross-correlation value. The maximum cross-correlation over a user defined range of allowable time shift is then used as a basis for determining the three components of the cross-correlation rating. For calculating the phase rating, if the model signal was shifted less than a user defined minimum, the rating receives a score of 1. If the curve is shifted more than a specified maximum, the score is zero. For phase shifts between the specified minimum and maximum, the score is determined based on a regression relationship. Following the time shift, the magnitude rating is computed by comparing the square of the areas between the curves and the time axis. The final magnitude rating is then determined as a ratio between the two areas raised to a user defined exponent. Lastly, the shape rating of the signal is calculated using the maximum cross-correlation value.

ISO Metric

Similar to CORA, the goal of the ISO metric was to combine a number of different rating metrics to robustly evaluate the correlation between two signals. Initially, the ISO established technical committee evaluated the CORA corridor technique and the Error Assessment of Response Time Histories (EARTH) [32] techniques to combine a corridor and cross correlation rating. Ultimately, the committee established an overall metric based on the CORA corridor algorithm, and an updated version of the EARTH score referred to as the Enhanced EARTH metric (EEARTH) [31].

Similar to the CORA cross-correlation metric, the total EEARTH rating is built on the individual phase, magnitude, and shape components. However, while the general components of the EEARTH metric are similar to CORA, there are unique features within the algorithms that differentiate the two. The phase metric of the EEARTH rating is used to assess phase lag between the model and test curves. Using a pre-defined maximum allowable percentage time-shift, the model curve is iteratively shifted left with discrete time step intervals and the cross-correlation between the truncated curves is calculated. Next, the test curve is shifted left over discrete time step intervals and the same calculation is performed. If the time shift is greater than or equal to the maximum allowable time shift, the score is 0. If the maximum cross-correlation value occurs with no time shift, the score is 1. For time shifts in between these values, the rating is calculated using a regression method [11,31]. The time shifted and truncated curves are then used to calculate the magnitude score.

Similar to CORA, the EEARTH magnitude rating measures differences in amplitude between the two curves. However, the EEARTH magnitude rating applies a dynamic time warping (DTW) algorithm prior to measuring discrepancies between the signals. The function of DTW is to expand and compress the time axis to align key components of the curve (such as local maxima and minima). This is all based on minimizing a local cost function [31]. Once the curves have been shifted, truncated, and warped, the magnitude error is calculated as a ratio of the difference in amplitude between the two signals based on a vector norm calculation. If the difference between the signals is less than the pre-defined threshold, the magnitude rating is 1. If the amplitude difference is greater than the maximum allowable magnitude error threshold, the score is 0. For values in between, the score is calculated using a regression function.

Lastly, the shape rating is calculated based on the test curve and the shifted, truncated model curve with no DTW applied. The two curves are divided into time intervals and the average slope is calculated at each interval. The shape score is then determined by calculating the ratio of the difference in slope between the model and test curves to the test curve. If there is no difference between the model and test curve, the shape score is 1. If the difference exceeds a pre-define threshold, the score is 0. Values in between are calculated based on a regression function.

Application of OE Methods

The kinetic and kinematic time-history traces obtained from the models were run through CORA v3.5 and ISO. When applying CORA, suggested default values were used for all parameter controls except for the phase range. For evaluating the phase shift, the allowable time shift range was changed from 3 to 12 percent to a range of 5 to 15 percent. For the application of ISO, all recommended weights and parameters set forth by the standard were applied. A total of 58 time history traces were obtained from the simulations. This provided a diverse sample of signals allowing for a robust comparison of the techniques. Because both techniques produce a phase, magnitude, and shape score, 174 (58 x 3) pair-wise comparisons were made. The corridor score of each technique was not included in the pair-wise analysis because the underlying algorithm is the same for both CORA and ISO. For each component of the cross-correlation rating metric, the comparative scores were cross-plotted and used to evaluate correlations. In this analysis, the coefficient of determination, R^2, was used to highlight differences in the two techniques. Statistical tests for significant differences between the two were also determined using a Wilcoxon matched-pairs signed rank test and a significance value of $\alpha = 0.05$.

Survey of Subject Matter Experts

The survey component of this study was approved by the Wake Forest School of Medicine's Institutional Review Board (IRB #39944). To evaluate how the OE techniques compare to real world interpretation, a survey was distributed to subject matter experts (SMEs) to obtain a scoring of how well the computational traces match the experiments in terms of phase, magnitude and shape. The objective was to compare the SME based scores to the quantitative results obtained from CORA and ISO to determine which, if either, is more in line with SME assessment. Participants were asked to complete a one-time, electronic survey designed to provide their interpretation of how well a subsample of 15 time history traces from the full dataset compared to experimental traces. Participation was limited to individuals with training or expertise in computational modeling and model validation techniques. Participants were contacted for inclusion in the study via email and all responses were anonymized prior to analysis. Demographic information including work title, affiliation, years of experience in biomechanics, and years of experience in model validation/signal analysis were requested. In order to answer the research question, the survey was sent to 69 SMEs with an expected participation of 50%. The sample size for the survey was determined by calculating the minimum number of survey questions and participants needed to detect a Cronbach's alpha of 0.9 assuming type 1 error of 0.05, a two-sided test, and a 80% power. Prior to evaluating the curves, participants were introduced to the terminology used in the study to ensure a reasonable baseline. However, no coaching was conducted in order to ensure that participants were not led to focus on specific curve attributes for evaluation. Each participant was asked to rank the phase, magnitude, and shape for the subsample of 15 curves, presented in a randomized order, on a scale of 0-100 that could be directly mapped to the scale implemented in CORA and ISO. The results were then analyzed using a 1-sample t-test that tested whether the mean from the sample was the same as CORA or ISO independently for phase, shape, or magnitude.

RESULTS

All simulations normally terminated without numerical error. In each case, simulations were visually inspected for localized areas of instability and were found to be stable. To illustrate the impacts evaluated in this study, a time lapse of each simulation can be seen in Figure 1.

The time history signals for each impact condition were exported and compared to the experimental data using both CORA and ISO. Because of the variety of data obtained from these simulations, the OE methods were evaluated using signals that ranged from good to poor correlation.

The overall average scores for the two algorithms, including the corridor metrics, were 0.60 and 0.56 for CORA and ISO respectively. No signal-based weighting approach was used for off-axis signals which may produce low correlation scores, but are also low magnitude compared to the resultant (e.g. shear vs. normal loading). Overall scores typically emphasize the dominant signals [33,34]. For both techniques, the corridor and cross-correlation scores were computed with the recommended weight factors (see Equations (1) and (2) where Z_{corr} stands for corridor score and $Z_{CrossCor}$ stands for cross-correlation score). With regards to magnitude, CORA rated the curves with an average score of 0.49 ± 0.27, whereas ISO rated the signals lower in general with an average rating of 0.38 ± 0.36. For phase, the average CORA score was 0.72 ± 0.38 and the average ISO score was 0.69 ± 0.30. Lastly, shape scores were rated as 0.71 ± 0.34 in CORA and 0.61 ± 0.16 in ISO. Overall, each of the components of the ISO cross-correlation score were lower on average compared to CORA, indicating a stricter rating of the signals.

$$CORA\ Score = 0.5 * Z_{Corr} + 0.5 * Z_{CrossCor} \quad (1)$$
$$ISO\ Score = 0.4 * Z_{Corr} + 0.6 * Z_{CrossCor} \quad (2)$$

Cross-plots for each component metric of CORA and ISO can be seen in Figure 2. In these plots, the respective magnitude, phase, and shape scores were aggregated from each simulation and compared using linear regression. As such, each point on the plot represents the CORA and ISO scores for a single time history trace. The phase scores for each metric

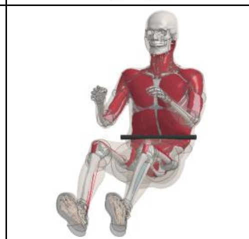
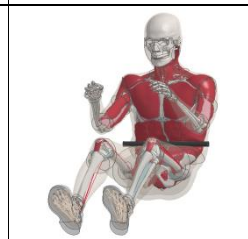
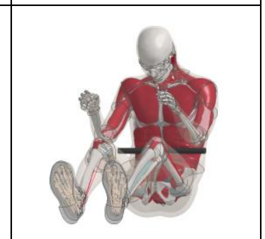
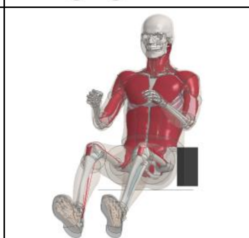
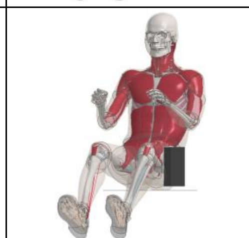
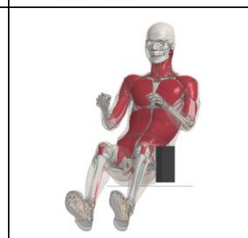
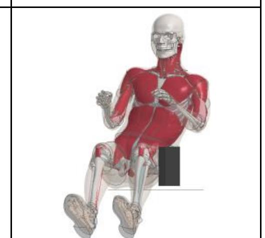
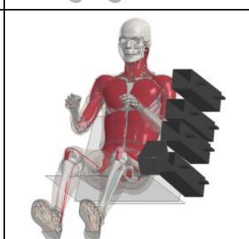
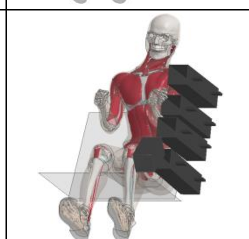
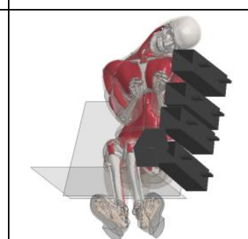
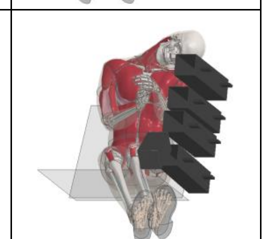
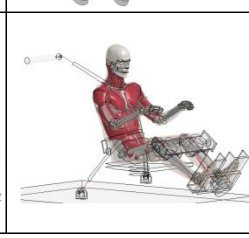
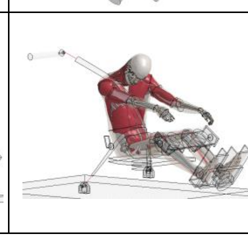
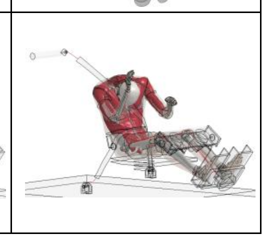
Impact Type	T = 0 ms	T = 1/3	T = 2/3	$T = T_{final}$
Thoraco-Abdominal Impact				
Frontal Abdomen Impact				
Lateral Pelvis Impact				
Lateral Sled				
Frontal Sled Test				

Figure 1. Simulation time-lapse of the M50 for each impact condition

were found to have the strongest correlation with an R^2 value of 0.66. Shape scores were found to have the weakest correlation with an R^2 value of 0.27. With regards to statistical comparison, the differences between CORA and ISO were found to be statistically significant for each component rating metric with p values of 0.003, 0.002, and 0.016 for phase, magnitude, and shape respectively.

Survey Responses

In total, 40 responses were collected from the survey solicitation. Participants were primarily from academia (72%), followed by industry (15%) and government (13%). More than 33% of participants had 10+ years of biomechanics and signal analysis experience, with more than 60% having 5+ years of experience.

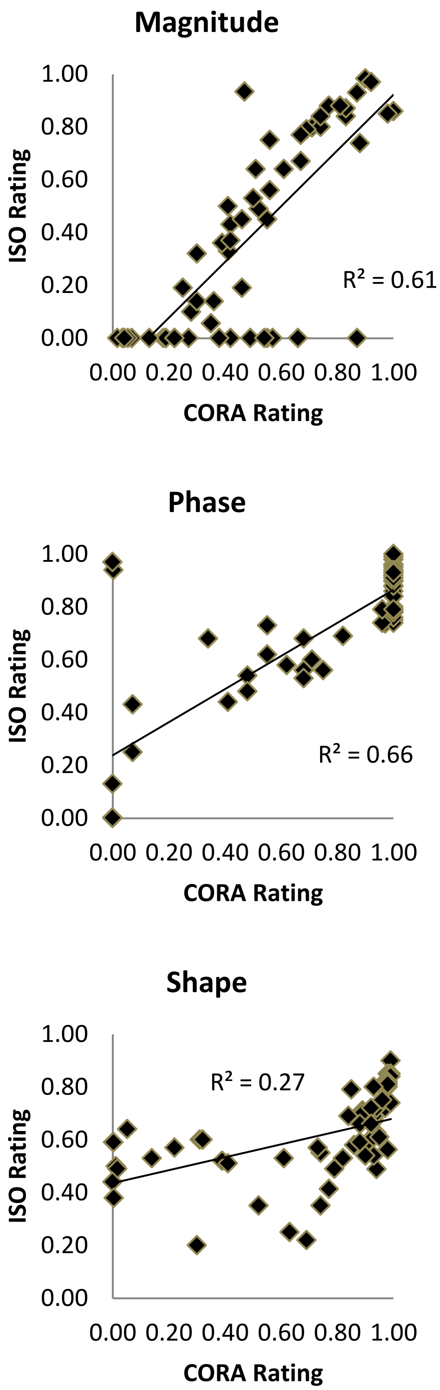


Figure 2. Correlation analysis of each component rating metric

The average response for the phase, magnitude, and shape characteristics for each of the 15 curves can be seen in Figure 3. In Figure 3, the bars represent the average of all survey responses for a particular curve and rating metric. Overall, volunteers rated the magnitude scores lowest with an average score of 0.52 across all 15 curves. The magnitude rating also had the largest variation in responses with respect to the average with a coefficient of variation of 0.38, indicating the widest variation in SME assessment. The phase rating was given the highest scores with an average of 0.70. The phase rating also had the lowest average of coefficient of variation (c_v = 0.28), indicating the greatest agreement in SME assessment.

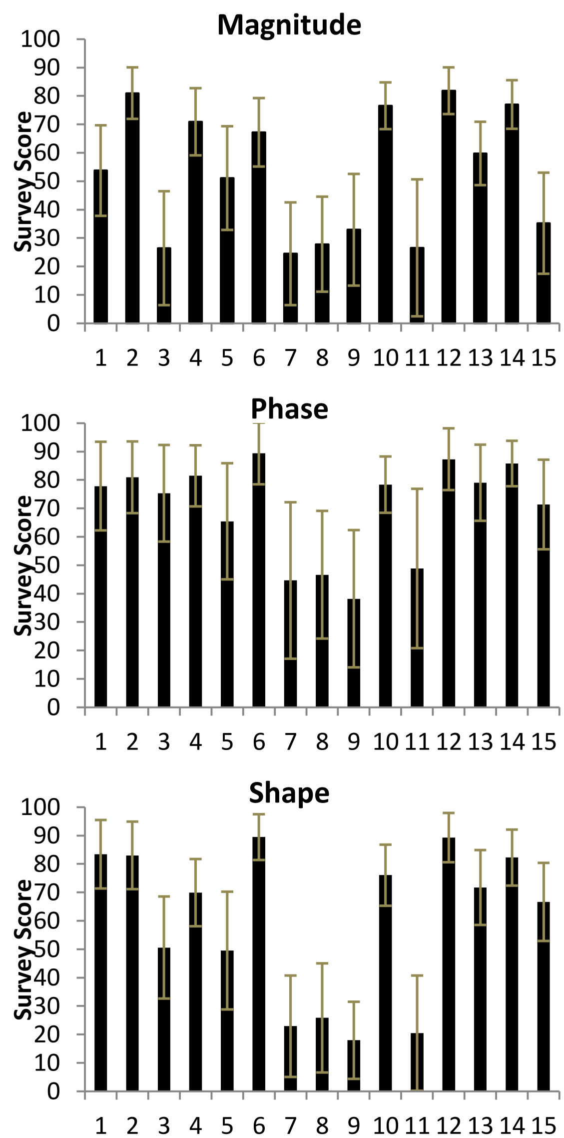


Figure 3. Survey responses for each curve

When comparing the SME responses to the ISO standard, the null hypothesis that the SME responses and ISO results were the same was rejected for magnitude ($p<0.001$), but was not rejected for the shape and phase metrics (p-values of 0.79 and 0.10 respectively). With regards to CORA, the null hypothesis was rejected for the shape and phase metrics (p-values of 0.005 and <0.001 respectively), but was not rejected for the magnitude rating ($p = 0.79$). From a real world perspective, this indicates that SME responses agreed with the ISO interpretation for phase and shape, but did not agree with the ISO magnitude rating. Conversely, the SME responses agreed with the CORA interpretation for magnitude, but not the phase and shape ratings.

For further evaluation, the percent difference in peak value between the model and experimental curves was compared to the average volunteer magnitude rating for each of the 15 curves in the survey. The percent difference in peaks was moderately correlated to the magnitude ratings with a Spearman's rho of -0.58. The average magnitude ratings were also compared to the percent difference in area under the curve for each of the curves evaluated in the survey. In this case, area under the curve showed strong correlation to the magnitude ratings with a Spearman's rho of -0.79.

DISCUSSION

As objective evaluation techniques are increasingly applied to the validation of computational human surrogate models, it is important to investigate how variations in different rating metrics can affect the overall estimation of model validity. The goal of this study was to apply both the CORA and ISO objective rating metrics to the same set of data derived from simulations of the GHBMC M50-O finite element model. While we strive for objective evaluations, we also note that the totality of how an engineer may view a signal is beyond what can be encapsulated in three nominally orthogonal measures (magnitude, phase and shape). Thus in this work we sought to find which algorithms are more likely to be in agreement with evaluations made by experts in the field. As such, the CORA and ISO interpretations were also compared to real world interpretations from SMEs.

The cross-plots depicted in Figure 2 show general trends for each component metric. For example, phase scores for both ISO and CORA had higher scores on average compared to the magnitude and shape ratings. However, more interesting are cases where one technique assigns a score of nearly 1 to a curve, and the other technique assigns the same curve a score closer to 0. Using these techniques as they were intended, this means the user is to interpret that one technique says the signal is a good match to the experimental data, but the other technique says the model does not represent the real world test.

This indicates that both techniques have limitations, and therefore must be combined with engineering judgment prior to drawing a final conclusion regarding a model's validity. However, as the survey responses showed, overall interpretation of the shape and phase response from SMEs tended to agree with the outputs from the ISO technique. For magnitude, the CORA method tended to more closely agree with the SMEs. This indicates that using the area of the signal may be a more intuitive means of assessing magnitude ratings. This finding also agrees with the strong correlation between the SME magnitude ratings and the percent difference in area for the curves in the survey. However, as the CORA magnitude rating uses a squared area ratio to assign a score, differences in polarity between the experimental and model curves can lead to artificially high scores. Therefore, when comparing signals with flipped polarity, this limitation in the CORA magnitude rating should be addressed in the future by including a polarity correction factor.

Overall, the CORA magnitude rating and ISO phase and shape ratings were found to provide the most intuitive scores when comparing model and experimental curves (Table 1). However, both techniques tend to give higher scores on average to the phase rating compared to the other component rating metrics. This can lead to biased total scores when phase is equally weighted with magnitude and shape. A similar trend was seen in SME interpretations, with an average phase score that was 30% and 17% higher than the subsampled magnitude and shape scores respectively. Therefore, in some applications, it may be necessary for the user to increase the exponent governing phase scores between 0 and 1 to make the regression equation either quadratic or cubic. Also, either isolated or in combination, the interval over which phase shift is permitted could be reduced to more strictly govern the phase rating. In general, this would more strictly govern the overall rating and enable researchers to discern potentially required model updates, such as viscoelastic adjustments.

Table 1.
Summary of agreement for OE techniques and SME interpretation

Component Rating Metric	OE Technique in Agreement with SMEs	p-value
Magnitude	CORA	0.79
Phase	ISO	0.1
Shape	ISO	0.79

When reporting final quantitative evaluations for validation, researchers commonly report the average of each component rating metric to give a global view of the model response. However, it may be necessary to discriminate between signals of varying magnitude to give a clearer picture of model behavior. For example, orthogonal signals (responses on the x, y, and z axes) often have motion on a primary axis (ex. X-axis) or primary plane of motion (ex. X-Y plane). In these cases, there are one or two off-axis responses that do not have the same scale as the primary motion. Therefore, in certain applications it may be appropriate to apply a weighting calculation to apply more weight to the scores of plots with greater magnitude. Davis et al. proposed an approach to weight objective evaluations based on a weighting factor derived from the peak values of the experimental mean traces [34].

CONCLUSIONS

Determining how closely a model matches an experiment is paramount for modelers. The goal of the OE methods evaluated in this work is to replace the subjectivity inherent in this process with a numerical score, yet it is clear from the results that ostensibly objective methods can produce different interpretations for the same data. This study provides a framework to critically compare results from each method, and highlights the relative strengths and weaknesses of each. In addition, a survey of SMEs allowed for the OE outputs to be compared to real world interpretation of model performance.

On average, ISO produced lower ratings than CORA, indicating a stricter evaluation of the model performance. The comparison also indicated statistically significant differences between the two techniques for each component rating metric, both in terms of the direct comparison between ISO and CORA and the comparison to SME interpretation.

Ultimately, the findings of the study suggest that using a mixed approach to reporting objective ratings, using the magnitude method in CORA and the ISO shape and phase methods, may be the most intuitive method to analyze model performance. However, it is noted that there are limitations for considering a model validated based solely on the outputs of OE techniques. While the OE methods evaluated in this study provide valuable insight with regards to model response, all OE analysis should be performed in conjunction with engineering judgment and other practical considerations. These include the ability to match signal peaks, which are often used by biomechanists as a correlate for the overall response or injury risk, and what experimental data is available to calculate the factors used in the techniques.

REFERENCES

[1] Tabiei A. and Wu J. "Roadmap for crashworthiness finite element simulation of roadside safety structures." Finite Elements in Analysis and Design, *2000*, vol. 34: pp. 145-157.

[2] Davis M. L., Vavalle N. A. and Gayzik F. S. "An evaluation of mass-normalization using 50th and 95th percentile human body finite element models in frontal crash." in *International Research Council on Biomechanics of Injury*, Lyon, France, 2015.

[3] Untaroiu C. D., Crandall J. R., Takahashi Y., Okamoto M., Ito O. and Fredriksson R. "Analysis of running child pedestrians impacted by a vehicle using rigid-body models and optimization techniques." Safety science, *2010*, vol. 48: pp. 259-267.

[4] Gaewsky J. P., Weaver A. A., Koya B. and Stitzel J. D. "Driver injury risk variability in finite element reconstructions of Crash Injury Research and Engineering Network (CIREN) frontal motor vehicle crashes." Traffic injury prevention, *2015*, vol. 16: pp. S124-S131.

[5] 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." Ann Biomed Eng, *2013*, vol. 41: pp. 497-512.

[6] Vavalle N. A. "Validation of the Global Human Body Models Consortium Mid-Sized Male Model in Lateral Impacts and Sled Tests." Wake Forest University, 2012.

[7] Sprague M. A. and Geers T. L. "A Spectral-Element Method for Modelling Cavitation in Transient Fluid-Structure Interaction." Int J Numer Meth Engng, *2004*, pp. 2467-2499.

[8] Rhule H. H., Maltese M. R., Donnelly B. R., Eppinger R. H., Brunner J. K. and Bolte J. H. "Development of a New Biofidelity Ranking System for Anthropomorphic Test Devices." Stapp Car Crash J, *2002*, vol. 46: pp. 477-512.

[9] 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*, vol. 14: pp. S87-S94.

[10] Gehre C., Gades H. and Wernicke P. "Objective Rating of Signals Using Test and Simulation Responses." in *The 21st ESV Conference*, Stuttgart, Germany, 2009.

[11] Barbat S., Fu Y., Zhan Z., Yang R. J. and Gehre C. "Objective Rating Metric for Dynamic Systems." in *Enhanced Safety of Vehicles* Seoul, Republic of Korea, 2013.
[12] Poulard D., Subit D., Nie B., Donlon J.-P. and Kent R. W. "The contribution of pre-impact posture on restrained occupant finite element model response in frontal impact." Traffic injury prevention, *2015,* vol. 16: pp. S87-S95.
[13] Gehre C. and Stahlschmidt S. "Assessment of dummy models by using objective rating methods." in *22nd International Technical Conference on the Enhanced Safety of Vehicles, Washington, DC*, 2011.
[14] Gayzik F. S., Moreno D. M., Geer C. P., Wuertzer S. D., Martin R. S. and Stitzel J. D. "Development of a Full Body CAD Dataset for Computational Modeling: A Multi-Modality Approach." Annals of Biomedical Engineering, *2011,* vol. 39: pp. 2568-2583.
[15] Li Z., Kindig M. W., et al. "Rib Fractures Under Anterior-Posterior Dynamic Loads: Experimental and Finite-Element Study." Journal of Biomechanics, *2010,* vol. 43: p. 228.234.
[16] Soni A. and Beillas P. "Modelling hollow organs for impact conditions: a simplified case study." Computer methods in biomechanics and biomedical engineering, *2015,* vol. 18: pp. 730-739.
[17] Shin J., Yue N. and Untaroiu C. D. "A finite element model of the foot and ankle for automotive impact applications." Ann Biomed Eng, *2012,* vol. 40: pp. 2519-31.
[18] DeWit J. A. and Cronin D. S. "Cervical Spine Segment Finite Element Model for Traumatic Injury Prediction." Journal of the Mechanical Behavior of Biomedical Materials, *2012,* vol. 10: pp. 138-150.
[19] Yang K. H., Hu J., White N. A., King A. I., Chou C. C. and Prasad P. "Development of Numerical Models for Injury Biomechanics Research: A Review of 50 Years of Publications in the Stapp Car Crash Conference." Stapp Car Crash J, *2006,* vol. 50: pp. 429-90.
[20] Toyota. "Documentation of Total Human Model for Safety (THUMS) AM50 Pedestrian/Occupant Model." Toyota Motor Corporation November 2010.
[21] Hayes A. R., Vavalle N. A., Moreno D. P., Stitzel J. D. and Gayzik F. S. "Validation of simulated chestband data in frontal and lateral loading using a human body finite element model." Traffic Inj Prev, *2014,* vol. 15: pp. 181-6.
[22] GHBMC. "GHBMC M50 Users Manual ", 2011.
[23] Viano D. C. "Biomechanical Responses and Injuries in Blunt Lateral Impact." Stapp Car Crash J, *1989,* vol. 33: pp. 113-142.
[24] Hardy W. N., Schneider L. W. and Rouhana S. W. "Abdominal impact response to rigid-bar, seatbelt, and airbag loading." Stapp Car Crash J, *2001,* vol. 45: pp. 1-32.
[25] Bouquet R., Ramet, M., Bermond, F., Caire, Y., Talantikite, Y., Robin, S., Voiglio, E. "Pelvis human response to lateral impact." in *16th International Technical Conference on the Enhanced Safetfy of Vehicles*, Windsor, ON, Canada, 1998, pp. 1665-1686.
[26] Cavanaugh J. M., Walilko T. J., Malhotra A., Zhu Y. and King A. I. "Biomechanical Response and Injury Tolerance of the Pelvis in Twelve Sled Side Impacts." SAE Technical Paper, *1990,* vol. 902305: p. 16.
[27] Cavanaugh J. M., Zhu Y., Huang Y. and King A. I. "Injury and Response of the Thorax in Side Impact Cadaveric Tests." SAE Technical Paper, *1993,* vol. 933127: p. 23.
[28] Shaw G., Parent D., et al. "Impact Response of Restrained PMHS in Frontal Sled Tests: Skeletal Deformation Patterns Under Seat Belt Loading." Stapp Car Crash J, *2009,* vol. 53: pp. 1-48.
[29] Ash J., Shaw C. G., Lessley D. and Crandall J. "PMHS Restraint and Support Surface Forces in Simulated Frontal Crashes." in *2012 JSAE Annual Congress (Spring)*, Pacifico Yokohama, 2012.
[30] Thunert C. "CORA Release 3.6 User’s Manual." GNS mbH, Germany, *2012,*
[31] (ISO) I. O. f. S. "Road vehicles - Objective rating metric for non-ambiguous signals." in *ISO/TS 18571*, 2013.
[32] Sarin H., Barbat S., Yang R. J., Kokkolaras M., Hulbert G. and Papalambros P. "Comparing Time Histories for Validation of Simulation Models: Error Measures and Metrics." Journal of Dynamic Systems, Measurement, and Control, *2010,* vol. 132: pp. 061401-061401.
[33] Pietsch H. A., Bosch K. E., et al. "Evaluation of WIAMan technology demonstrator biofidelity relative to sub-injurious PMHS response in simulated under-body blast events." Stapp car crash journal, *2016,* vol. 60: pp. 199-246.
[34] Davis M. L., Koya B., Schap J. M. and Gayzik F. S. "Development and Full Body Validation of a 5th Percentile Female Finite Element Model." Stapp Car Crash J, *2016,* vol. 60: pp. 509-544.