Development of a Robust Human Body Model Qualification Methodology: Ensuring Biofidelity for Occupant Protection Assessments in Virtual Testing

Felix Ressi, Bernd Schneider, Desiree Kofler, Vishnuvel Kannan, Devon C. Hartlen, Duane S. Cronin, Corina Klug

Abstract Virtual testing using human body models (HBMs) is becoming increasingly important in vehicle safety assessment. However, unlike anthropomorphic test devices, HBMs lack standardised qualification methods for harmonised occupant simulations. This study presents a novel methodology for the objective qualification of HBMs in virtual testing. A set of HBM-neutral validation setups was developed, covering impactor, tabletop, and sled tests under controlled boundary conditions. These setups, along with a standardised scoring methodology, enable the evaluation of different HBMs while accounting for inherent variability in post-mortem human subject (PMHS) test data. A key outcome of this research is the basis for Technical Bulletin CP 550, which European New Car Assessment Programme (Euro NCAP) will implement from 2026, marking a significant step toward standardised HBM qualification. The study focuses on frontal and far side impact scenarios, with the methodology extendable to additional load cases and body regions. The publicly available assessment framework ensures reproducibility and facilitates continuous model improvement. This work lays the foundation for enhanced biofidelity in virtual testing, paving the way for broader HBM integration into consumer rating protocols.

Keywords Biofidelity, human body models, model qualification, occupant protection, virtual testing.

I. INTRODUCTION

Virtual methods play a crucial role in modern vehicle development. Since 2024, they have also been used in European New Car Assessment Programme (Euro NCAP) consumer information testing. In addition to physical crash tests, the far side assessment now incorporates simulation data using the World Side Impact Dummy (WorldSID) anthropomorphic test device (ATD) [1].

In research and development, human body models (HBMs) are increasingly employed to evaluate occupant restraint systems and vulnerable road user (VRU) safety measures. However, HBMs are not yet included in regulatory requirements or consumer information ratings. The primary challenge is the lack of a standardised methodology for qualifying HBMs in virtual testing (VT), which currently only exists for pedestrian head impact time assessment [2]. While all state-of-the-art HBMs are validated during development and updates, the respective load cases and modelling approaches may vary considerably between the models.

Unlike virtual models of ATDs, which can be directly validated against their physical counterparts, for example, by applying an ISO rating [3] to time-history curves, HBMs require a different approach. Since HBMs have no direct physical standardised counterpart available for injurious testing, their validation typically relies on post-mortem human subject (PMHS) tests. However, PMHS tests involve real human subjects who are not standardised, leading to greater variability than ATD tests. While the ISO rating [3] is established within Euro NCAP's VT scheme [4], it is designed to compare two signals with one another and does not take variability of test data into account. Furthermore, it cannot be used with signals exhibiting hysteresis, e.g. force-deflection curves. An effective HBM qualification methodology however, must account for the inherent variability in PMHS test results and should be applicable to all types of signals.

This study contributes to the development of a procedure to qualify HBMs for virtual testing. The procedure aims to ensure that HBMs produce biomechanical responses representative of human behaviour, within the variability observed in corresponding PMHS test data, so that they deliver comparable results for fair and consistent virtual assessments. It is designed to be applicable by simulation engineers with limited biomechanical

F. Ressi (+43 316 873 30363, felix.ressi@tugraz.at) is a Project Senior Scientist, Bernd Schneider a Researcher and Corina Klug is Associate Professor at the Vehicle Safety Institute at Graz University of Technology in Austria. D.C. Hartlen is a Research Associate and D.S. Cronin is a Professor at Canada Research Chair in Trauma Biomechanics and Injury Protection at the University of Waterloo, Ontario, Canada.

expertise. While the methodology is generally independent of the type of assessment (e.g., adult occupants, children, VRUs), this study focused on average male occupant models used to predict kinematics in frontal and far side impacts as a representative starting point. The study evaluates the performance of HBMs that are either widely used in vehicle development or have recently been developed to address current limitations. This evaluation aims to establish a baseline for future refinement of the qualification procedure and to support harmonisation efforts across HBM development.

II. METHODS

The complete qualification methodology, applicable to HBMs used within Euro NCAP protocols, consists of requirements in four key areas: anthropometry (i.e., geometric requirements), model quality, HBM outputs, and biofidelity in validation setups. While all requirements must be fulfilled to qualify an HBM for VT, the present study focuses on the biofidelity assessment in particular. Hence, the next sections describe these steps in detail.

Preparation of HBM-neutral Simulation Setups

Within a group of international experts, potential validation setups were collected and evaluated based on the following criteria: level of certainty on boundary conditions (BCs), experience with the setup, previous problems explored with the BCs, relevance for far side assessment, relevance for frontal assessment, relevance for full-body kinematics, relevance for tissue-based rib fracture assessment, availability and accessibility of setups (setups with intellectual property (IP) issues were not included). The individual ratings provided by participants were consolidated into a group rating, with all criteria weighted equally. Setups were selected if they were rated as having medium to high relevance and did not meet any exclusion criteria, such as intellectual property concerns or previously identified issues with boundary conditions. This resulted in a final group of 19 validation setups for HBM qualification, including nine impactor-style setups [5–13], five tabletop-style setups [14–18] and five sled-style setups [19–24]. Multiple load cases were selected for each validation setup, such as varying loading rate, loading angle, or impactor geometry. At least one response of interest was extracted from each validation setup and load case to assess HBM performance. These responses were selected to capture different aspects of the HBM response. The final set of validation setups, load cases, and responses of interest is provided in Table A IV.

To maximise reproducibility and minimise user-dependent variations in setting up the simulations, the input decks were designed in an HBM-neutral manner. This included systematic instructions for HBM positioning based on anatomical landmarks, well-defined sets for contact definitions, and guidelines for configuring additional outputs (e.g., displacement time-histories of individual vertebrae) where necessary. With the exception of tabletop-style setups, the spine was constrained during the settling phase to maintain the HBMs as close as possible to their originally developed and validated state. Apart from maximising reproducibility and facilitating model setup for users with limited biomechanical expertise, these measures enable future implementation of simulation model traceability procedures. Such procedures, e.g. hashing, can ensure that the submitted results were obtained with unaltered boundary condition models and that the HBM was not changed between qualification and the simulations performed in virtual testing itself [25].

Although the setups were initially created for LS-Dyna, they were designed for easy conversion to other solver codes. For instance, to ensure compatibility with Engineering System International Group Virtual Performance Solution (ESI Group VPS), parameter names were restricted to six characters.

In addition to the simulation models, Python assessment scripts in Jupyter notebook format [26] were provided for direct data extraction and visualisation from simulation outputs. The simulation models were shared with HBM developers, who integrated their HBMs, ran the simulations, and executed the assessment scripts. The resulting comma separated values (CSV) files were shared with the authors of this study for analysis.

As part of the assessment, the HBM response was evaluated using the scoring methodology described in the following section.

Scoring Methodology

The scoring methodology objectively rated the fit of simulation data to experimental PMHS data using statistical response corridors generated by ARCGen [27]. ARCGen uses arc-length re-parameterisation and signal registration to compute an average response and corridors from experimental data that accounts for variability in both axes. ARCGen can be applied across the range of responses used for model assessment, including responses that may not terminate at a common coordinate in either axes (like force-displacement data), may be

highly oscillatory (like time-series kinematic data), or may be hysteretic (like load-unload data from impact testing). For a comprehensive coverage of the ARCGen methodology, the reader is referred to the original source [27], while a concise synopsis is provided herein for informational purposes. Initially, each signal is scaled based on a mean bounding box, and the scaled signals are then re-parameterised using their normalised arc-length. To calculate the mean bounding box of all input signals, the maximum and minimum values in the x and y directions are extracted from each input signal. The mean bounding box is calculated by taking the mean of minimum and maximum values in both directions across all input signals. Scaling signals by the mean bounding box helps ensure that differences in magnitude between the x and y axes do not affect the arc-length re-parameterisation process.

Once re-parameterised with respect to normalised arc-length, input signals undergo the process of signal registration, with the objective of aligning critical features with respect to normalised arc-length to prevent distortion of the computed average and corridors. To perform registration, warping functions are introduced for each signal. The signal registration process has a key control parameter, the number of warping control points (*nWarpCtrlPts*), which ensures accurate alignment of features. The number of warping control points is a user-defined parameter that must be selected in accordance with the shape of the input signals. In the third and final step of the ARCGen process, the mean and standard deviation in the x and y directions are calculated for each point with respect to normalised arc-length. As variability is assessed in two dimensions, the combined x and y standard deviations represent the major and minor axes of an elliptical region around each point in the average response that represents variability in the input signals. The average response and ellipses form a key part of two rating methods developed for this study to assess HBM performance: the Ellipse Score and Dynamic Arc-Length Warping (DALW). For both rating methods, averages and standard deviations were computed using consistent parameters, defined in Table A I.

The first rating method, the Ellipse Score, is a measure of how well the simulation data matches the average response of PMHS data while accounting for experimental variability. To calculate the Ellipse Score, the simulation response, average response and standard deviations were scaled by the mean bounding box used to compute the average response and corridors with ARCGen, followed by uniform resampling to 500 points with respect to normalised arc-length. The goodness-of-fit of the scaled and resampled simulation response to the scaled and resampled average response was assessed in a point-wise manner. The approach is visualised in Fig. 1 and described below. At each resampled point, a score was computed based on how closely the point of the simulation response p_i fell to the average response c_i. If the simulated point fell within an ellipse e_i^{inner} (centred at the corresponding point on the average response) representing one standard deviation of variability, a score of 1 was given for that point. If the simulated point fell outside an ellipse e_i^{outer} of two standard deviations, a score of 0 was given for that point. If the simulated point fell between the two ellipses, a score between 0 and 1 was given based on how closely the simulated point fell to the inner ellipse. To assign this score, a line was projected from the average point to the simulated point, intercepting both ellipses. l_1 was defined as the distance of the simulated point to the intersection of the line with the inner ellipse and l_2 was defined as the distance of the simulated point to the intersection of the line with the outer ellipse. The score was then calculated as $\frac{l_2}{l_1+l_2}$ for this case. The three cases and their respective scores are shown in Fig. 1. The Ellipse Score was finally determined as the mean of all the individual scores.

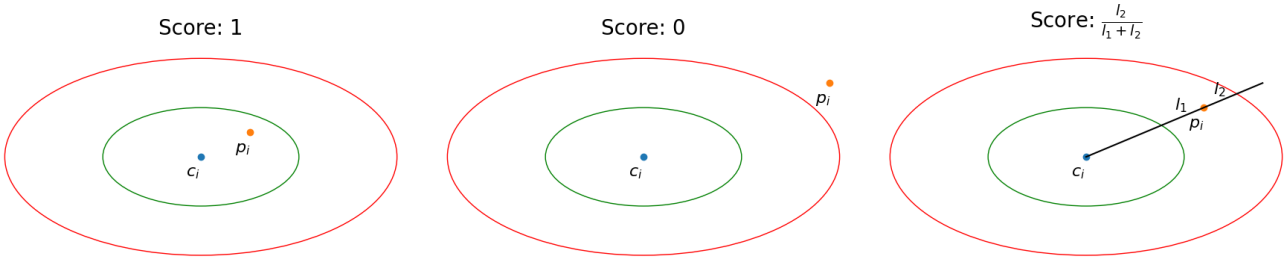

Fig. 1. Score is one if the point p_i is inside the inner ellipse e_i^{inner} shown in green. Score is zero if the point p_i is outside the outer ellipse e_i^{outer} shown in red. Score is $\frac{l_2}{l_1+l_2}$ if the point p_i is between the two ellipses.

The stringent mapping of each point p_i to the corresponding ellipses, e_i^{inner} and e_i^{outer}, can result in cases where the simulation data appears close to the average response but still receives a low score. Additionally, it

renders the score susceptible to minor alterations in the simulation results, e.g. noise or small differences in initial position, which change the arc-length of the response. To address these challenges, a second score, termed Dynamic Arc-Length Warping (DALW) Score, was developed. The DALW Score extends the described Ellipse Score with dynamic time warping (DTW). DTW is a similarity measure for sequences that has been introduced originally for speech recognition [28]. A general description of the DTW algorithm can be found in [28] and [29]. For the DTW algorithm, a global constraint (windowing) and a local constraint (step pattern) can be set. These constraints were chosen to be the same as the ISO 18571:2024 standard [3], since the ISO standard is well established and aimed at vehicle safety applications [3]. The constraints are repeated here for reference. For the global constraint, a Sakoe-Chiba window was used where the window size was 10% of the length of the data. This means that the window size was 50 for the DALW score calculation. Using notation consistent with [3], the step pattern was chosen as

$$d_{tw}[i,j] = d(i,j) + \min(d_{tw}[i-1,j], d_{tw}[i,j-1], d_{tw}[i-1,j-1]) \quad (1)$$

where $d_{tw}[i,j]$ is an entry of the cumulative cost matrix and $d(i,j)$ is an entry of the local cost matrix. The calculation of the entries of the local cost matrix $d(i,j)$ for the DALW score differs from [3]. Whereas DTW generally uses the squared distances between the corresponding points to populate the local cost matrix, the DALW score uses a similar rating approach as the Ellipse Score. $d(i,j)$ is calculated as $1 - s(i,j)$ where $s(i,j)$ is the score of point p_i with the ellipses e_j^{inner} and e_j^{outer}, as presented above. Therefore, if the point p_i lies within the inner ellipse e_j^{inner}, the corresponding entry is 0. If the point p_i lies outside the outer ellipse e_j^{outer}, the entry is 1. If the point lies between the two ellipses, the entry is between 0 and 1. Having calculated the optimal warping path with the DTW algorithm, the DALW Score is finally calculated as $1 - \frac{DTW_{opt}}{k_{opt}}$ where DTW_{opt} is the total cost of the optimal warping path and k_{opt} the number of indices of the optimal warping path. This essentially calculates the mean of all scores $s(i,j)$ along the optimal warping path. It is important to note that the DALW Score and the Ellipse Score are the same if the optimal warping path would match the diagonal of the local cost matrix.

Qualification Procedure

Eventually, the scores of eight HBMs that are either widely used in vehicle development or have been developed very recently were derived to act as a first baseline for defining thresholds that are achievable in the short term. All criteria laid out in the procedure (including model anthropometry, model quality, and a list of required model outputs and the minimum thresholds for the two scores in all validation setups) must be fulfilled before a model can be used for VT.

III. RESULTS

Preparation of HBM-neutral Simulation Setups

All HBM-neutral setups are publically available on OpenVT [30] for LS-Dyna and VPS. However, all results presented in this study are exclusively based on simulation results from the LS-Dyna setups. As the aim of this study was not to compare HBM performance using the developed qualification methodology but to discuss the methodology itself, all HBM results are provided in anonymised form.

TABLE I
OVERVIEW ON HBMS UTILISED IN THE PRESENT STUDY

Model	Version
GHBMC M50-0	v6.2
GHBMC M50-0S	v2.3
HANS 50M	v1.0
HBM-C 50M	v1.5
SAFER HBM	v11.1
THUMS 50M	v4.1
THUMS 50M	v7
VIVA+ 50M	V1.1

Specifically, results from eight HBMs comprising 13 result sets, submitted by eleven different organisations, were available. Table I lists the HBMs in alphabetical order and their respective model versions. Not all organisations submitted results for all load cases from the most recent validation setup versions, which limited the data available for deriving the thresholds based on the scoring. However, complete data were available for four different HBMs, including three complete sets of results for one of them, resulting in six complete result sets overall.

Scoring Methodology

For all available responses of interest from the submitted HBM results, the Ellipse Score and DALW Score were calculated. The individual scores for each HBM in each response of interest from each validation setup and load case are provided in the Appendix.

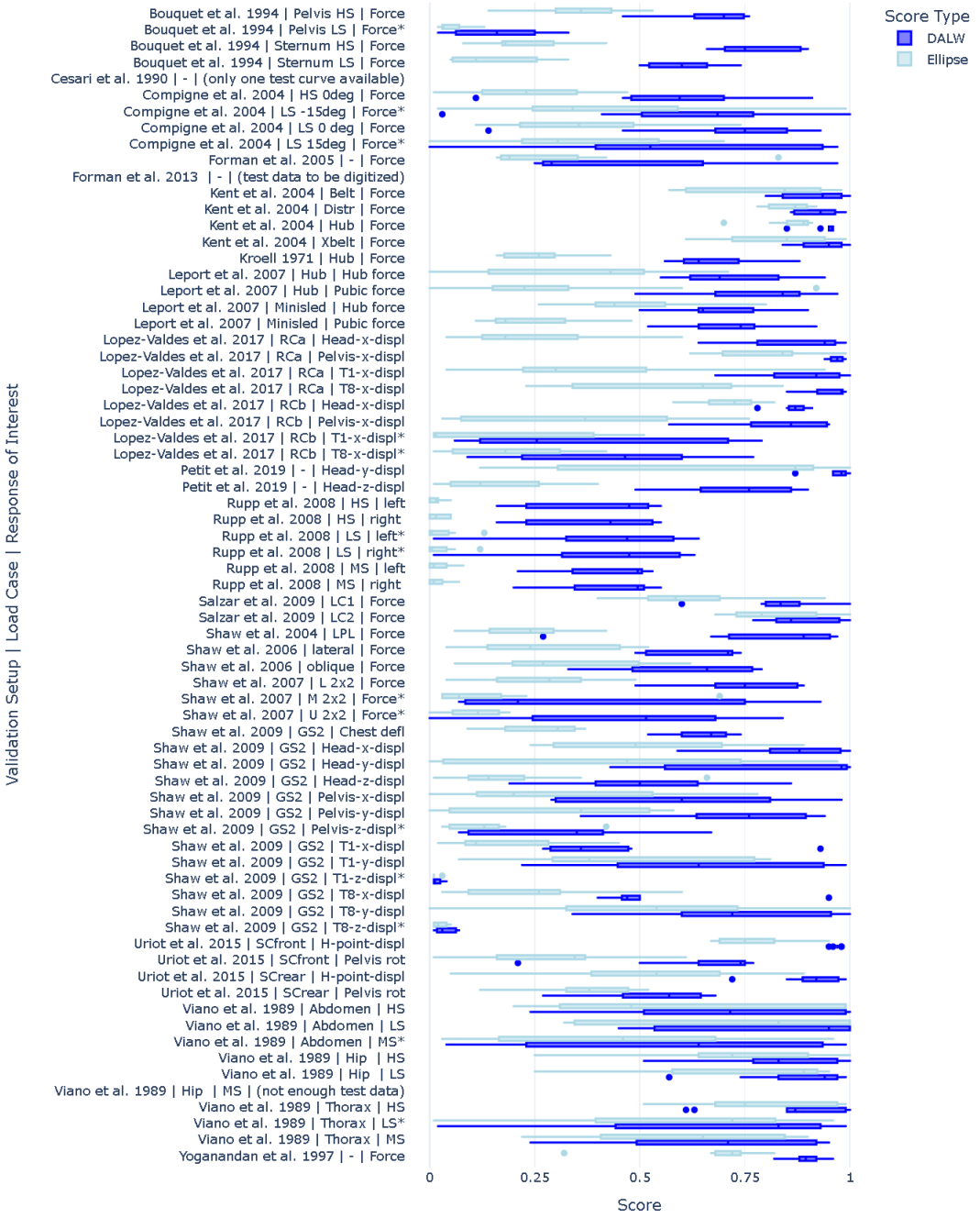

Fig. 2. Boxplots illustrating the Ellipse Score (light blue) and the DALW Score (dark blue) for all load case configurations and their evaluated signals in alphabetical order. For responses marked with an asterisk, no thresholds for Ellipse and DALW Scores were set, in effect removing them from the qualification requirements at the time of writing this study. For the three load cases for which the score calculation was not performed, additional information is provided in parentheses.

The boxplots in Fig. 2 provide an overview on the range of values for the two scores obtained in each evaluated signal for all load cases. For load cases with fewer than three experimental data traces, the score calculation was not performed.

Looking at the 14 responses currently not included in the qualification (marked with an asterisk in Fig. 2), two patterns emerge: One group yields low values in both, Ellipse and DALW Score, for instance the force response in the low speed pelvis impact from Bouquet et al. 1994 (Pelvis LS I Force) or the displacement in Z direction for the pelvis (Pelvis-z-displ), the first thoracic vertebra (T1-z-displ) and the eight thoracic vertebra (T8-z-displ) from Shaw et al. 2009. The other generally group exhibits larger ranges of scores, particularly for the DALW Score, for instance the force-compression response in the low speed thorax impact from Viano 1989.

Furthermore, Fig. 2 illustrates that scores above 0.9 (up to 1) are achieved more frequently using the DALW Score compared to the Ellipse Score (25% and 7% of all scores respectively). Furthermore, they also show that scores below 0.1 (down to 0) occur more commonly for the Ellipse Score than the DALW Score (23% and 6% respectively).

To better clarify the relationship between Ellipse and DALW Score, Fig. 3 shows a scatter plot of Ellipse Score on the horizontal, and DALW Score on the vertical axis. Three responses of interest from three load cases were selected to avoid overloading the plot. Specifically, the force deflection curve of a tabletop belt loading test [18], the impactor force from a lateral minisled test to the hip [11], and the head displacement in X direction for a frontal sled test [20] were selected. A scatter plot for all responses of interest from all load cases is provided in the Appendix. Based on the algorithm behind the Dynamic Arc-Length Warping, the DALW Score is equal to the Ellipse Score at worst. Fig. 3 visualises that while there are cases that produce similar scores for both metrics (i.e., points within ±10% of the 1:1-ratio line) particularly towards the extremes, there are also cases that yield considerably higher DALW Scores than Ellipse Scores. For easier comparison, Fig. 4 shows the scores for these three load cases in the same boxplot style as the overview in Fig. 2.

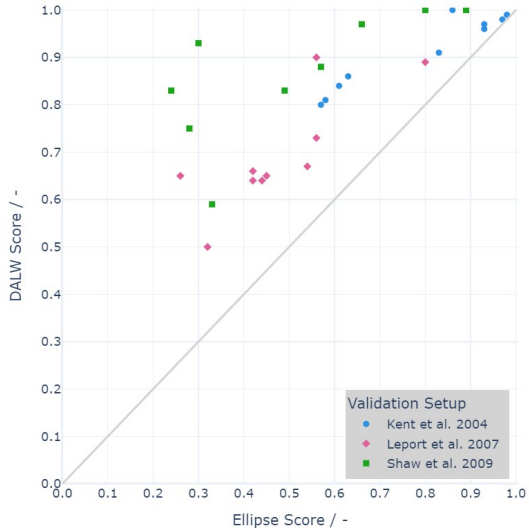

Fig. 3. Scatter plot illustrating the relationship between the Ellipse Score and DALW Score for three selected responses of interest from load cases.

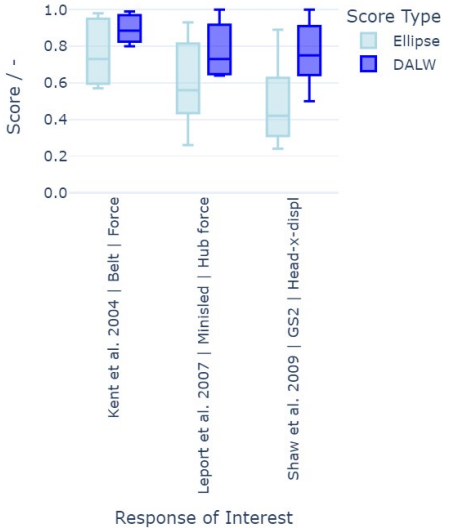

Fig. 4. Boxplots for the scores of three selected responses of interest from load cases.

Fig. 5, Fig. 6, and Fig. 7 provide examples of HBM results compared to the ARCGen characteristic average and corridors (i.e., envelopes based on the 1 SD and 2 SD ellipses), using the same three evaluations as above.

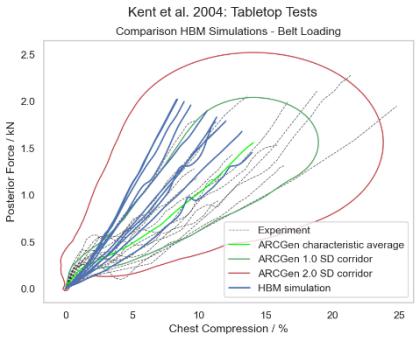

Fig. 5. Force-displacement curves from a tabletop belt pull test [18] with the characteristic average, the ARCGen corridors (1 SD, 2 SD) and the HBM simulation results.

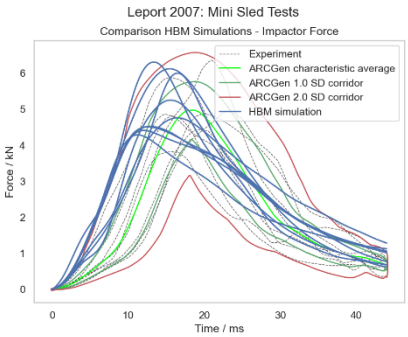

Fig. 6. Force time-history curves from a hip impact [11] with the characteristic average, the ARCGen corridors (1 SD, 2 SD) and the HBM simulation results.

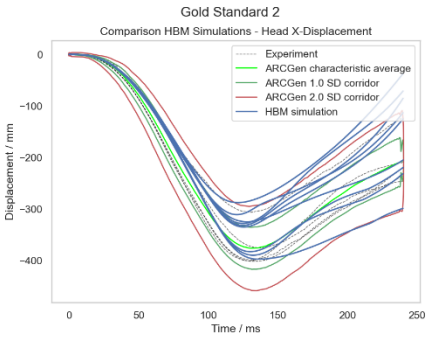

Fig. 7. Head X displacement time-history curves from a Gold Standard 2 test [20] with the characteristic average, the ARCGen corridors (1 SD, 2 SD) and the HBM simulation results.

For the tabletop belt pull test [18] in Fig. 5, all HBM force-deflection curves fall within the envelope of the 2 SD ellipses (with the exception of one curve, which is just below until approx. 2.5% chest compression). While this does not necessarily mean that each point of the HBM curves along the arc-length is also within the respective 2 SD ellipse of the ARCGen characteristic average curve, the lowest resulting Ellipse Score value (as shown in Fig. 3) of 0.57 suggests that this is mostly the case. While most HBM curves appear slightly stiffer than the characteristic average, two are close to it. Furthermore, all but three curves (which are all from the same HBM) are within the range of the individual PMHS curves shown in the plot.

The force time-history curves from a hip impact [11] in Fig. 6 show similar trends for all HBM curves initially. In the first 10 ms to 20 ms, the HBMs' behaviour is stiffer than the ARCGen characteristic average. At around 10 ms, three of the HBM curves (which are all from the same HBM) reach their maximum with the force reducing more gradually compared to the characteristic average. Two HBM curves show maximum force levels very similar to, and three considerably higher ones than the characteristic average. All force peaks occur earlier than that in the characteristic average curve. Nevertheless, apart from the initial stiffness, the HBM results cover the range of PMHS responses, which is also reflected in the scores summarised in the middle boxplot in Fig. 4.

The head X displacement time-history curves in Fig. 7 show that the HBM results fall within the 2 SD ellipse envelope, apart from three curves. One enters rebound early, at approx. 120 ms, and immediately leaves the upper 2 SD ellipse envelope. Another curve falls outside the 2 SD envelope at approx. 150 ms, after exhibiting a faster rebound motion compared to the other models. The third curve enters rebound latest, at approx. 130 ms, and exhibits a slower rebound movement, subsequently leaving the 1 SD ellipse envelope at approx. 155 ms and running along the edge of the 2 SD envelope from 200 ms until the end. The first two of these three curves also resulted in two of the lowest Ellipse Scores in this response of interest (0.33 and 0.28, with the third curve yielding a value of 0.57 respectively). For all head X displacement time-history curves shown in Fig. 7, the scores increase considerable when calculating the DALW Score (0.59, 0.75, and 0.88 respectively).

Qualification Procedure

Based on the calculated scores for the submitted HBM results, the thresholds for qualification were set to the minimum score achieved by these models. Simulation results that seemed implausible compared to the other HBMs or did not fulfil the quality criteria were not considered for the determination of this threshold value. The most common examples for results that were deemed implausible included signals that were considerably shorter and signals where no reliable determination of initial contact was possible. In addition, the scoring was not applied to cases where the threshold would have been below 0.1. This was determined for Ellipse and DALW Score individually. The rationale behind this decision was that below this, there are considerable differences between the reference curve and the simulation response and setting limits this low does not ensure biofidelic behaviour anyway. Furthermore, it showed to be very sensitive within this very low range.

Two potential paths for model qualification are available for HBM users. New HBMs, i.e., a model that has not been qualified before, need to perform the *full qualification*. Specifically, it needs to pass the requirements on

anthropometry, model quality, and model outputs, as well as the minimum thresholds for the two scores in all validation setups. For HBMs, for which this full qualification has been performed successfully (i.e., HBMs which have been approved for virtual testing), users can opt for a simplified qualification. For this *comparability demonstration*, only two validation load cases need to be run. They consist of a tabletop belt-loading test [18] and a frontal sled test [20]. These load cases were selected as the scores of the HBMs were reasonable high for these signals in these load cases and to demonstrate that the HBM has not been modified (and that the simulation environment the HBM is used in, performs as expected). Specifically, the resulting scores in these load cases are compared to the reference value documented in the report for the full qualification. They need to be within a tolerance of ±0.1 of the Ellipse Score and DALW Score. In case this tolerance is exceeded, the user can still opt to perform the full qualification and run all validation setups and use the model in virtual testing if all scores pass the thresholds.

IV. DISCUSSION

During the development of the HBM-neutral simulation setups, several load cases with insufficient test data (i.e., less than three PMHS responses) were identified (also see Table A IV in the appendix). This was the case for two load cases, specifically a tabletop setup [15], and one specific impact speed for a lateral hip impact setup [12]. For these two setups, the lack of data means they cannot be used for biofidelity scoring and are only suitable for comparative assessments between HBMs. For one farside sled setup [19] no raw data, but only corridors, were available, which is why the developed approach was not applicable. Whilst writing this manuscript, efforts were underway to obtain the original test data of the farside sled setup, which would enable application of the scoring methodology in future.

One goal of the HBM-neutral setups was to ensure a high degree of reproducibility. Specifically, the results for any given HBM should be very similar, regardless of who sets up the simulation. This is particularly important when considering that even running identical simulations on different numbers of CPUs or with different model decomposition approaches has been shown to influence results [31]. An example for one HBM is shown in Fig. 8.

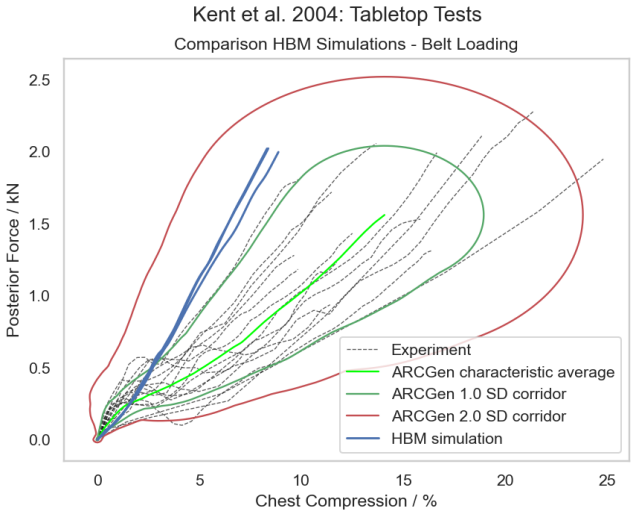

Fig. 8. Comparison of the response of one selected HBM in a tabletop belt-loading test [18], set up by three different organisations, indicating that reasonable reproducibility is achieved by the HBM-neutral simulation setups.

Of the three HBM responses, two were almost identical, represented by the slightly thicker blue line in Fig. 8. For these two, both scores (Ellipse and DALW Score) were within 0.01 of one another. The third HBM response was slightly less stiff and yielded scores deviating between 0.05 and 0.06 from the other two. Thus, regardless of which one would be the basis for comparison, all three would have fulfilled the ±0.1 criterion applied for the two comparability load cases (with this tabletop belt-loading test [18] being one of the two). While this indicates that the setups work as intended and that reasonable reproducibility is maintained, more results, ideally from a systematic study involving a large number of HBMs applied by users with different backgrounds setting several load cases are needed to analyse the influence of the setups and the user on the resulting variability in more detail.

Overall, the two developed scoring metrics showed quite different behaviour. In general, an ideal scoring method would fulfil the following criteria: The variation of test data should be incorporated into the score calculation. The method should be applicable to all kinds of data, such as data with monotonic or non-monotonic X values. The method should be sensitive to physical differences in test and simulation data, for example differences in energy. The method should be robust and not sensitive to normal scatter in simulation results, e.g. due to different numbers of central processing units (CPUs) used to run the simulation [31]. While the Ellipse Score satisfies most of these criteria, it can be overly sensitive to small deviations in simulation results. In particular, it may penalise cases where the simulation response initially deviates from the test corridor but re-enters it shortly afterwards, despite an overall good agreement. The development of the DALW score was driven by these limitations. While it is more robust than the Ellipse Score, it appears to be overly forgiving in some cases. It is worth noting, however, that as shown for the example in Fig. 7, the DALW Score more clearly differentiated between the curve that is most different visually (i.e., the curve entering rebound and leaving the 2 SD envelope earliest, at approx. 120 ms) and the curves that were within the range of PMHS responses for longer periods of time. An even more extreme example for a curve with low Ellipse Score but high DALW Score, is represented by the green square point on the upper left in Fig. 3. The HBM response resulted in an Ellipse Score of only 0.3, although it never ran outside the 2 SD envelope (the respective curve in Fig. 7 mostly followed the upper 1 SD envelope and ended as the fourth curve from the top). While the Ellipse Score would suggest it is comparable to the three curves discussed above, its DALW Score of 0.93 seems more fitting for a curve that is very close to one of the PMHS responses. This is in agreement with engineering intuition. However, since neither scoring metric was found to be clearly superior, all results for both scores are provided in this study.

Several parameters (see also Table A I in the Appendix) control the behaviour of ARCGen. Among them, signal registration and the parameter *nWarpCtrlPts* has the greatest potential to impact the ellipse calculation depending on shape of the underlying signals. According to [27], the number of prominent inflection points shared among the input signals is generally an appropriate value for this *nWarpCtrlPts*. Given the variability in test response shapes, a sensitivity study was conducted to assess the effect of the number of warping control points on the scoring outcome. Five responses from five selected load cases were analysed. Three of these, already used to illustrate differences between the Ellipse and DALW Scores in Fig. 3 to Fig. 7, included the force deflection curve of a tabletop belt-loading test [18], the impactor force from a lateral minisled test to the hip [11], and head displacement in the X-direction during a frontal sled test [20]). Two additional responses were included: the force deflection curve from a lateral abdominal impact [12] and the head displacement in the Y-direction for a far side sled test [24]. Together, these responses cover a diverse range of impact locations, loading directions, and signal types, providing a robust basis for evaluating the sensitivity of the scoring to the *nWarpCtrlPts* parameter. In this study, the number of control points was varied from zero to five across these load cases using results from five HBMs for which data were available in all cases. The results of the parameter study are shown in Fig. A 2 in the Appendix. These results demonstrate that the effect of signal registration (and the *nWarpCtrlPts* parameter) is relatively minor as the number of warping control points increases. However, it is also important to note that scores were generally poorer if signal registration was not performed on the input signals (setting *nWarpCtrlPts* to 0). Given relatively low impact of warping control points on score, *nWarpCtrlPts* was set to one for all validation cases.

Limitations

In qualification procedures, a balance must be struck between scientific rigour and practical usability. The individuals performing qualifications for vehicle rating purposes are often not the original HBM developers or experts in biomechanics. This is one reason why certain pragmatic simplifications were introduced in the load case evaluations, in contrast to what is typically seen in scientific publications. However, these simplifications must not compromise the correct representation of boundary conditions. Of course, the correct representation of boundary conditions is not only relevant with respect to any modelling simplifications, but needs to be ensured generally.

To enable this first step towards virtual testing with HBMs, while continuing to improve boundary condition models that did not produce satisfactory results based on the submitted data and user feedback, the thresholds for the Ellipse and DALW scores were removed for 14 responses of interest across seven load cases (see Table A IV). Some of these responses of interest showed consistently low scores across all submitted HBM results, suggesting a systematic difference between the PMHS tests and the simulations. Others exhibited wider

interquartile ranges compared to similar responses. While this might indicate that the qualification procedure is capable of distinguishing more biofidelic models, a closer examination of the results, combined with user feedback, showed that specific modelling choices affected the performance of different HBMs in different ways. Although it would have been possible to remove these load cases entirely, the decision was taken to retain them. This allows results from future qualifications to support further model improvements, helping to ensure that simulation models more accurately represent the corresponding PMHS tests. In summary, care was taken to avoid unphysical simulation responses, and the ongoing refinement of the setups is encouraged, particularly in response to user feedback and the increasing volume of submitted data. While all simulation setups may be improved in the future, those currently used in the qualification procedure were considered to adequately represent the boundary conditions of the PMHS tests for this initial step towards virtual testing with HBMs.

It is important to note that the score thresholds were not defined in a way that would incentivise excessive tuning of the HBMs. Before any tightening of these thresholds, a further review of all validation setups will be carried out. This review will not only examine the suitability of the simulation setups themselves, but also re-evaluate the quality and representativeness of the underlying PMHS data. In particular, attention will be given to whether the test subjects reflect the intended target population for the HBMs. For example, PMHS series with a predominance of very fragile individuals may not be appropriate if the HBMs are intended to represent a broader or more average population. Nonetheless, the inclusion of models that are still undergoing continuous improvement represents a limitation of the present study.

Another aspect of the scoring methodology concerns the specific portion of each response that is used for scoring. The relevant start and end points of each signal are provided for all responses in Table A IV in the Appendix (columns x_min and x_max). For the force-displacement responses shown in Fig. 5, the force data were restricted to values equal to or greater than 5% of the maximum force. This filtering was applied individually to each PMHS and HBM response to maintain consistency. This approach was agreed upon following discussions with several stakeholders. However, it is important to acknowledge that the data could be processed in other valid ways, which might result in differences in scoring outcomes, particularly regarding calculated stiffness. As such, the choice of signal range and filtering method represents a methodological limitation of the present study, and may influence the comparability of results depending on the specific implementation.

A large number of validation setups were included in the qualification procedure to ensure the robustness of HBM biofidelity. However, the resulting volume of simulations also creates an administrative burden, as the results submitted for each HBM must be reviewed by the respective testing organisation. Furthermore, while it is important to separate calibration and validation datasets to ensure robust and predictive model behaviour, the number of available PMHS test setups is limited. As a result, using many of them in a qualification procedure reduces the pool of setups available for model calibration.

To address this, the number of validation setups may be reduced in the future. This would lessen the burden of both generating and reviewing HBM results, while potentially freeing up more load cases for calibration. Nevertheless, the validation setups described in the present study are designed with future traceability in mind and are inherently different from the in-house calibration setups used by HBM developers. In developer-specific setups, the HBM can be pre-positioned using individual PMHS landmark data and even morphed to match PMHS anthropometry. By contrast, the presented HBM-neutral setups require the HBM to start each simulation in its default posture.

This standardisation allows relevant parts of the boundary condition model, control cards, and the HBM itself to be hashed, ensuring the integrity of the simulation results [25]. Accordingly, the criteria defined in the present study should be understood as a final verification of model performance, and not as targets for calibration.

Following the initial introduction of the qualification procedure, the minimum thresholds required to pass will be reviewed. In load cases where it is deemed beneficial, thresholds may be raised to encourage further improvement of the HBMs. Such adjustments will be made only after a critical evaluation of the validation setups to ensure that boundary conditions are reliably captured by the simulation models, thereby supporting increased HBM biofidelity.

V. CONCLUSION

This study is the first to directly compare the behaviour of contemporary HBMs using a set of shared simulation models and introduces a concept for the objective qualification of HBMs for virtual testing. The HBM-neutral

validation setups and the assessment routines, including the scoring methodology developed in this study, are publicly available. This enables the evaluation of current and future HBMs under consistent boundary conditions.

Although the present work focuses on frontal and far-side load cases and thoracic injury assessment, the proposed methodology can be extended to other body regions and loading scenarios, provided that appropriate validation cases are developed.

The results form the technical foundation of Euro NCAP's Technical Bulletin CP 550, which will be implemented from 2026 onwards. This represents an important first step towards a standardised qualification procedure for HBMs. Continuous refinement of the methodology is encouraged as more experimental data become available.

VI. ACKNOWLEDGEMENTS

The authors thank all contributors who provided simulation models of specific load cases, which served as the basis for the HBM-neutral setups. Where applicable, original developers are acknowledged on OpenVT. The authors also acknowledge the valuable feedback received during the development of the simulation setups, as well as the submitted HBM simulation results.

VII. REFERENCES

[1] Euro NCAP. (2023) Euro NCAP, *Far side occupant test & assessment procedure, Version 2.5*.

[2] Klug, C.; Feist, F.; Schneider, B.; Sinz, W.; Ellway, J.; van Ratingen, M. (2019): Development of a Certification Procedure for Numerical Pedestrian Models. International Technical Conference on the Enhanced Safety of Vehicles. 10-13 June. Eindhoven, Netherlands.

[3] ISO 18571. (2024): ISO 2024-05 - Road vehicles - Objective rating metric for non-ambiguous signals. ISO/TS 18571:2024(E),

[4] European New Car Assessment Programme (Euro NCAP). (2023): Virtual Far Side Simulation and Assessment Protocol v1.0.

[5] Bouquet, R.; Ramet, M.; Bermond, F.; Cesari, D. (1994): Thoracic and pelvis human response to impact. The 14th International Technical Conference on the Enhanced Safety of Vehicles,

[6] Compigne, S.; Caire, Y.; Quesnel, T.; Verries, J.-P. (2004): Non-Injurious and Injurious Impact Response of the Human Shoulder Three-Dimensional Analysis of Kinematics and Determination of Injury Threshold. 48th Stapp Car Crash Conference. NOV. 01, 2004.

[7] Kroell, C. K.; Schneider, D. C.; Nahum, A. M. (1971): Impact Tolerance and Response of the Human Thorax. 15th Stapp Car Crash Conference (1971). NOV. 17, 1971.

[8] Rupp, J. D.; Miller, C. S.; Reed, Matthew P.; Madura, N. H.; Klinich, K. D.; Schneider, L. W. (2008): Characterization of Knee-Thigh-Hip Response in Frontal Impacts Using Biomechanical Testing and Computational Simulations. Stapp Car Crash Journal, Vol. 52,

[9] Shaw, G.; Lessley, D.; Bolton, J.; Crandall, J. (2004): Assessment of the Thor and Hybrid III Crash Dummies: Steering Wheel Rim Impacts to the Upper Abdomen. SAE 2004 World Congress & Exhibition. MAR. 08, 2004.

[10] Shaw, J. M.; Herriott, R. G.; McFadden, J. D.; Donnelly, B. R.; Bolte, J. H. (2006): Oblique and lateral impact response of the PMHS thorax. Stapp Car Crash Journal, **50**: 147–167.

[11] Leport, T.; Baudrit, P.; Trosseille, X.; Petit, P.; Palisson, A.; Vallancien, G. (2007): Assessment of the Pubic Force as a Pelvic Injury Criterion in Side Impact. Stapp Car Crash Journal, Vol. 51: 467–488.

[12] Viano, D. C. (1989): Biomechanical Responses and Injuries in Blunt Lateral Impact. 33rd Stapp Car Crash Conference. 4.10.1989.

[13] Yoganandan, N.; Pintar, F. A.; Kumaresan, S.; Haffiner, M.; Kuppa, S. (1997): Impact biomechanics of the human thorax-abdomen complex. International Journal of Crashworthiness, **2**(2): 219–228.

[14] Shaw, G.; Lessley, D. et al. (2007): Quasi-static and dynamic thoracic loading tests: cadaveric torsos. IRCOBI Conference. 19.-21. September. Maastricht, The Netherlands.

[15] Cesari, D.; Bouquet, R. (1990): Behaviour of Human Surrogates Thorax under Belt Loading. Stapp Car Crash Conference. NOV. 05, 1990.

[16] Forman, J. L.; Kent, R.; Ali, T.; Crandall, J.; Bostrom, O.; Haland, Y. (2005): Biomechanical considerations for the optimization of an advanced restraint system: assessing the benefit of a second shoulder belt. IRCOBI Conference. 21.-23.9. 2005. Prague, Czech Republic.

[17] Salzar, R. S.; Bass, C. R.; Lessley, D.; Crandall, J. R.; Kent, R. W.; Bolton, J. R. (2009): Viscoelastic response of the thorax under dynamic belt loading. Traffic Injury Prevention, **10**(3): 290–296.

[18] Kent, R.; Lessley, D.; Sherwood, C. (2004): Thoracic Response to Dynamic, Non-impact Loading from a Hub, Distributed Belt, Diagonal Belt, and Double Diagonal Belts. 48th Stapp Car Crash Conference. November 1-3, 2004. Nashville, Tennessee.

[19] Forman, J. L.; Lopez-Valdes, F. et al. (2013): Occupant kinematics and shoulder belt retention in far-side lateral and oblique collisions: a parametric study. Stapp Car Crash Journal, **57**: 343–385.

[20] Shaw, G.; Parent, D. et al. (2009): Impact response of restrained PMHS in frontal sled tests. Stapp Car Crash Journal, **53**: 1–48.

[21] Lopez-Valdes, F. J. (2017) *Kinematic comparison between the THOR dummy, older volunteers and older PMHS in low-speed non-injurious frontal impacts*.

[22] Lopez-Valdes, F. J.; Mroz, K. et al. (2018): Chest injuries of elderly postmortem human surrogates (PMHSs) under seat belt and airbag loading in frontal sled impacts: Comparison to matching THOR tests. Traffic Injury Prevention, **19**(sup2): S55-S63.

[23] Uriot, J.; Potier, P. et al. (2015): Reference PMHS Sled Tests to Assess Submarining. Stapp Car Crash Journal, **59**: 203–223.

[24] Petit, P.; Trosseille, X. et al. (2019): Far Side Impact Injury Threshold Recommendations Based on 6 Paired WorldSID / Post-Mortem Human Subjects Tests. Stapp Car Crash Journal, **63**: 127–146.

[25] Edwards, M.; Forman, J. et al. (in press): Model Traceability Procedure for an Automotive-Safety Virtual-Testing Framework. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 10.-12.09. Vilnius, Lithuania.

[26] Granger, B. E.; Perez, F. (2021): Jupyter: Thinking and Storytelling With Code and Data. Computing in Science & Engineering, **23**(2): 7–14.

[27] 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**: 843148.

[28] Sakoe, H.; Chiba, S. (1978): Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, **26**(1): 43–49.

[29] Giorgino, T. (2009): Computing and Visualizing Dynamic Time Warping Alignments in R : The dtw Package. Journal of Statistical Software, **31**(7).

[30] Open Virtual Testing Organisation (OVTO). HBM4VT Validation Catalogue, https://openvt.eu/EuroNCAP/hbm4vt-validation-catalogue. [17.03.2025].

[31] Östh, J.; Pipkorn, B.; Forsberg, J.; Iraeus, J. (2021): Numerical Reproducibility of Human Body Model Crash Simulations. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 08.-10.09.2021. Munich, Germany.

VIII. APPENDIX

Parameters used for ARCGen

TABLE A I
PARAMETERS USED FOR ARCGEN

nWarpCtrlPts	NormalizeSignals	nResamplePoints	CorridorRes
1	On	500	500

Ellipse Score and DALW Score for HBMs 1 – 6 for all responses of interest

Please note that the HBMs are listed in randomised order.

TABLE A II
ELLIPSE SCORE (ES) AND DALW SCORE FOR HBMS 1 – 6 FOR ALL RESPONSES OF INTEREST

			HBM 1		HBM 2		HBM 3		HBM 4		HBM 5		HBM 6	
	Load case	Criteria	ES	DALW	ES	DALW	ES	DALW	ES	DALW	ES	DALW	ES	DALW
Bouquet et al. 1994	Pelvis LS	Force	0.03	0.25			0.02	0.02	0.03	0.2	0.03	0.14	0.09	0.16
	Pelvis HS	Force	0.3	0.75			0.14	0.5	0.36	0.74	0.25	0.46	0.53	0.72
	Sternum LS	Force	0.05	0.5			0.33	0.59	0.05	0.5	0.11	0.63	0.11	0.67
	Sternum HS	Force	0.18	0.74			0.42	0.75	0.18	0.7	0.37	0.76	0.08	0.86
Cesari et al. 1990	-	-												
Compigne et al. 2004	HS 0deg	Force	0.21	0.49	0.46	0.72	0.47	0.91	0.19	0.46	0.31	0.57	0.39	0.72
	LS 0 deg	Force	0.22	0.7	0.51	0.85	0.74	0.93	0.2	0.73	0.44	0.69	0.4	0.77
	LS 15deg	Force	0.23	0.45	0.54	0.94	0.57	0.93	0.23	0.42	0	0	0.7	0.96
	LS -15deg	Force	0.26	0.65	0.36	0.74	0.64	0.77	0.23	0.53	0.99	1	0.28	0.48
Forman et al. 2005	-	Force	0.17	0.27			0.3	0.47	0.17	0.27	0.83	0.97	0.42	0.74
Forman et al. 2013	-	-												
Kent et al. 2004	Hub	Force	0.9	0.96			0.81	0.93	0.9	0.96	0.88	0.95	0.88	0.95
	Belt	Force	0.58	0.81			0.93	0.96	0.57	0.8	0.83	0.91	0.98	0.99
	Xbelt	Force	0.71	0.89			0.92	0.96	0.73	0.91	0.72	0.84	0.99	1
	Distr	Force	0.81	0.86			0.86	0.94	0.8	0.86			0.89	0.96
Kroell 1971	Hub	Force	0.18	0.59	0.26	0.73	0.35	0.88	0.19	0.62	0.28	0.64	0.43	0.75
Leport et al. 2007	Hub	Hub force	0.45	0.7			0	0.57	0.46	0.68	0.29	0.55	0.14	0.83
	Hub	Pubic force	0.31	0.89			0	0.68	0.33	0.88	0.1	0.49	0.92	0.97
	Minisled	Hub force	0.45	0.65			0.56	0.9	0.44	0.64	0.56	0.73	0.26	0.65
	Minisled	Pubic force	0.18	0.74			0.11	0.78	0.18	0.76	0.28	0.61	0.48	0.65
Lopez-Valdes et al. 2017	RCa	Head-x-displ	0.35	0.94	0.08	0.79	0.6	0.99	0.36	0.96	0.04	0.64	0.14	0.94
		T1-x-displ	0.47	0.93	0.3	0.87	0.23	0.99	0.4	0.92	0.04	0.68	0.65	0.97
		T8-x-displ	0.69	0.98	0.71	0.94	0.37	0.87	0.65	0.98	0.23	0.85	0.84	0.99
		Pelvis-x-displ	0.62	0.94	0.93	0.99	0.78	0.98	0.84	0.97	0.69	0.94	0.84	0.96
	RCb	Head-x-displ	0.73	0.91	0.62	0.87	0.75	0.86	0.71	0.87			0.58	0.78
		T1-x-displ	0.42	0.76	0.01	0.09	0.01	0.06	0.36	0.66			0.01	0.17
		T8-x-displ	0.42	0.65	0.35	0.55	0.1	0.28	0.24	0.49			0.01	0.16
		Pelvis-x-displ	0.03	0.73	0.66	0.95	0.03	0.8	0.12	0.83			0.37	0.94
Petit et al. 2019	-	Head-y-displ	0.91	0.98			1	1	0.9	0.98	0.29	0.99	0.31	0.99
	-	Head-z-displ	0.25	0.9			0.12	0.83	0.29	0.76	0.02	0.49	0.06	0.6
Rupp et al. 2008	HS	left	0.02	0.52			0.01	0.23	0.01	0.54	0	0.17	0.05	0.49
		right	0.02	0.52			0.01	0.23	0.01	0.53	0	0.17	0.05	0.43
	MS	left	0	0.5			0.01	0.21	0	0.5			0.08	0.53

		right	0	0.5			0.01	0.2	0	0.5			0.07	0.55
	LS	left	0.03	0.6			0.13	0.55	0	0.31			0.01	0.34
		right	0.02	0.61			0.12	0.54	0	0.31			0.01	0.32
Salzar et al. 2009	LC1	Force	0.6	0.84			0.69	0.88	0.58	0.83	0.52	0.8	0.47	0.6
	LC2	Force	0.73	0.82			0.68	0.77	0.73	0.83			0.81	0.87
Shaw et al. 2004	LPL	Force	0.06	0.84			0.31	0.89	0.25	0.97	0.24	0.93	0.18	0.27
Shaw et al. 2006	lateral	Force	0.47	0.72			0.11	0.53	0.4	0.71	0.52	0.72	0.04	0.51
	oblique	Force	0.54	0.78			0.22	0.33	0.37	0.73	0.19	0.66	0.27	0.55
Shaw et al. 2007	U 2x2	Force	0.19	0.84			0.09	0.32	0.11	0.51				
	M 2x2	Force	0.03	0.08			0.1	0.2	0.04	0.09				
	L 2x2	Force	0.09	0.7			0.37	0.89	0.23	0.86				
Shaw et al. 2009	GS2	Head-x-displ	0.8	1	0.24	0.83	0.28	0.75	0.57	0.88			0.33	0.59
		Head-y-displ	0.04	0.58	0.67	0.99	0.47	1	0	0.43			0.95	1
		Head-z-displ	0.1	0.75	0.1	0.44	0.16	0.5	0.66	0.86			0.01	0.19
		T1-x-displ	0.27	0.47	0.09	0.28	0.24	0.36	0.45	0.93			0.11	0.31
		T1-y-displ	0.38	0.52	0.81	0.91	0.78	0.99	0.07	0.22			0.77	0.99
		T1-z-displ	0.01	0.01	0.01	0.01	**0.03**	0.04	0.01	0.01			0.01	0.02
		T8-x-displ	0.26	0.47	0.11	0.4	0.31	0.47	0.6	0.95			0.2	0.5
		T8-y-displ	0.49	0.68	0.89	0.94	0.22	0.34	0	0.42			1	1
		T8-z-displ	0.04	0.05	0.01	0.01	0.04	0.06	0.02	0.03			0.01	0.02
		Pelvis-x-displ	0.2	0.6	0.78	0.96	0	0.3	0.15	0.3			0.5	0.68
		Pelvis-y-displ	0.11	0.64	0.36	0.72	0.58	0.94	0	0.36			0.05	0.91
		Pelvis-z-displ	0.16	0.39	0.04	0.07	0.11	0.13	0.14	0.48			0.42	0.67
		Chest defl	0.35	0.74	0.34	0.72	0.15	0.69					0.29	0.67
Uriot et al. 2015	SCfront	H-point-displ	0.67	0.97			0.68	0.97	0.76	0.97	0.82	0.95	0.95	0.98
		Pelvis rot	0.37	0.74			0.16	0.64	0.35	0.75	0.36	0.75	0.3	0.74
	SCrear	H-point-displ	0.45	0.92			0.89	0.99	0.58	0.92			0.39	0.9
		Pelvis rot	0.37	0.52			0.46	0.62	0.38	0.57			0.52	0.68
Viano et al. 1989	Hip	HS	0.73	0.84			0.42	0.67	0.71	0.82	0.25	0.51	0.64	0.77
		MS												
	Abdomen	LS	0.89	0.94			0.25	0.57	0.89	0.94			0.54	0.74
		HS	0.99	0.99			0.49	0.72	0.99	0.99	0.44	0.67	0.31	0.51
		MS	0.96	0.99			0.26	0.32	0.66	0.91			0.15	0.2
		LS	1	1			0.56	0.78	1	1			0.33	0.46
	Thorax	HS	0.68	0.86			0.99	1	0.68	0.86	0.79	0.88	0.51	0.61
		MS	0.65	0.71			0.89	0.95	0.65	0.71			0.22	0.24
		LS	0.72	0.83			0.96	0.99	0.72	0.83			0.11	0.12
Yoganandan et al. 1997	-	Force	0.67	0.91			0.32	0.82	0.74	0.92	0.71	0.86	0.82	0.96

Ellipse Score and DALW Score for HBMs 7 – 13 for all responses of interest

Please note that the HBMs are listed in randomised order.

TABLE A III

ELLIPSE SCORE (ES) AND DALW SCORE FOR HBMS 7 – 13 FOR ALL RESPONSES OF INTEREST

			HBM 7		HBM 8		HBM 9		HBM 10		HBM 11		HBM 12		HBM 13	
	Load case	Criteria	ES	DALW	ES	DALW	ES	DALW	ES	DALW	ES	DALW	ES	DALW	ES	DALW
Bouquet et al. 1994	Pelvis LS	Force	0.03	0.25			0.03	0.06	0.08	0.33	0.03	0.05	0.13	0.3	0.04	0.07
	Pelvis HS	Force	0.31	0.75			0.44	0.63	0.37	0.7	0.3	0.67	0.41	0.76	0.47	0.63
	Sternum LS	Force	0.07	0.52			0.24	0.74	0.26	0.61	0.05	0.53	0.1	0.73	0.26	0.6
	Sternum HS	Force	0.18	0.66			0.16	0.89	0.25	0.71	0.17	0.7	0.18	0.9	0.31	0.9
Cesari et al. 1990	-	-														
Compigne et al. 2004	HS 0deg	Force	0.14	0.47			0.01	0.11	0.25	0.61	0.05	0.58	0.11	0.66	0.29	0.68
	LS 0 deg	Force	0.23	0.67			0.11	0.14	0.59	0.85	0.31	0.77	0.21	0.46	0.46	0.89
	LS 15deg	Force	0.21	0.37			0	0	0.28	0.48	0.38	0.67	0.33	0.57	0.55	0.97
	LS -15deg	Force	0.14	0.41			0.02	0.03	0.55	0.72	0.63	0.84	0.32	0.55	0.43	0.77
Forman et al. 2005	-	Force	0.16	0.25			0.27	0.43	0.17	0.28	0.16	0.25	0.37	0.71	0.19	0.29
Forman et al. 2013	-	-														
Kent et al. 2004	Hub	Force	0.9	0.96			0.85	0.96	0.7	0.85			0.9	0.96	0.91	0.96
	Belt	Force	0.63	0.86			0.97	0.98	0.61	0.84			0.93	0.97	0.86	1
	Xbelt	Force	0.78	0.94			0.97	0.98	0.61	0.84			0.94	0.97	0.92	1
	Distr	Force	0.87	0.93			0.92	0.98	0.89	0.93			0.92	0.99	0.78	0.87
Kroell 1971	Hub	Force	0.17	0.66	0.16	0.58	0.28	0.56	0.16	0.61	0.23	0.62	0.42	0.81	0.26	0.73
Leport et al. 2007	Hub	Hub force	0.41	0.62			0.51	0.94	0.56	0.75			0.71	0.87	0.06	0.68
	Hub	Pubic force	0.21	0.86			0.6	0.84	0.15	0.668			0.24	0.84	0.19	0.79
	Minisled	Hub force	0.42	0.64			0.8	0.89	0.32	0.5					0.42	0.66
	Minisled	Pubic force	0.21	0.77			0.45	0.92	0.12	0.52					0.17	0.72
Lopez-Valdes et al. 2017	RCa	Head-x-displ	0.18	0.86			0.18	0.75					0.19	0.98		
		T1-x-displ	0.2	0.83			0.28	0.79					0.94	1		
		T8-x-displ	0.25	0.98			0.63	0.95					0.74	0.99		
		Pelvis-x-displ	0.7	0.97			0.84	0.96					0.99	0.99		
		Head-x-displ	0.78	0.88			0.82	0.9					0.72	0.85		
	RCb	T1-x-displ	0.51	0.79			0.01	0.15					0.02	0.34		
		T8-x-displ	0.12	0.77			0.01	0.09					0.27	0.44		
		Pelvis-x-displ	0.47	0.89			0.76	0.95					0.37	0.57		
Petit et al. 2019	-	Head-y-displ	0.92	0.98			0.12	0.96					0.73	0.87	0.87	0.96
	-	Head-z-displ	0.21	0.89			0.01	0.7					0.4	0.66	0.07	0.85
Rupp et al. 2008	HS	left	0	0.55			0	0.5	0.05	0.44	0.02	0.16			0.02	0.46
		right	0	0.55			0	0.53	0.05	0.43	0.02	0.16			0.05	0.4
	MS	left	0	0.51			0.01	0.46			0.06	0.22			0.02	0.49
		right	0	0.52			0.01	0.48			0.05	0.21			0.01	0.49
	LS	left	0.06	0.64			0	0.56			0	0.39			0	0.01
		right	0.06	0.63			0	0.58			0	0.41			0	0.01
Salzar et al. 2009	LC1	Force	0.59	0.84			0.4	0.81	0.52	0.79			0.94	1	0.69	0.88
	LC2	Force	0.77	0.85			0.85	0.95					1	1	0.99	1

Shaw et al. 2004	LPL	Force	0.42	0.96	0.13	0.67								
Shaw et al. 2006	lateral	Force	0.22	0.49	0.24	0.74								
	oblique	Force	0.62	0.79	0.06	0.46								
Shaw et al. 2007	U 2x2	Force	0.12	0.52	0	0			0.14	0.68	0.19	0.68	0.02	0.17
	M 2x2	Force	0.03	0.6	0.11	0.22			0.03	0.07	0.23	0.9	0.69	0.93
	L 2x2	Force	0.04	0.49	0.49	0.89			0.32	0.66	0.35	0.74	0.25	0.76
Shaw et al. 2009	GS2	Head-x-displ	0.89	1	0.49	0.83					0.66	0.97	0.3	0.93
		Head-y-displ	0.04	0.56	0.01	0.56					0.97	0.99	0.6	0.98
		Head-z-displ	0.07	0.6	0.18	0.26					0.36	0.5	0.14	0.49
		T1-x-displ	0.32	0.47	0.09	0.29					0.02	0.48	0.07	0.27
		T1-y-displ	0.3	0.44	0.35	0.45					0.27	0.64	0.72	0.92
		T1-z-displ	0.01	0.01	0.01	0.04					0.01	0.02	0.01	0.01
		T8-x-displ	0.31	0.48	0.31	0.47					0.03	0.5	0.04	0.42
		T8-y-displ	0.36	0.66	0.58	0.97					0.54	0.72	0.68	0.95
		T8-z-displ	0.05	0.07	0.04	0.07					0.01	0.01	0.02	0.02
		Pelvis-x-displ	0.19	0.31	0	0.29					0.25	0.76	0.62	0.98
		Pelvis-y-displ	0.04	0.62	0.53	0.89					0.52	0.81	0.47	0.76
		Pelvis-z-displ	0.13	0.35	0.18	0.38					0.05	0.1	0.03	0.07
		Chest defl	0.32	0.67	0.37	0.54					0.21	0.52	0.09	0.66
Uriot et al. 2015	SCfront	H-point-displ	0.72	0.97	0.82	0.97			0.77	0.96	0.74	0.98	0.69	0.97
		Pelvis rot	0.34	0.68	0.38	0.77			0.61	0.75	0.15	0.21	0.01	0.5
	SCrear	H-point-displ	0.54	0.9	0.05	0.72			0.87	0.98	0.63	0.97	0.37	0.85
		Pelvis rot	0.31	0.46	0.45	0.64			0.51	0.66	0.12	0.27	0.33	0.46
Viano et al. 1989	Hip	HS	0.9	0.97	1	1	0.82	0.92			0.68	0.8	1	1
		MS												
		LS	0.92	0.97	0.59	0.86	0.93	0.97			0.87	0.96	0.95	0.99
	Abdomen	HS	0.99	1	0.27	0.49	0.47	0.71			0.2	0.24	0.99	0.99
		MS	0.74	0.98	0.03	0.04	0.46	0.64			0.17	0.24	0.62	0.92
		LS	0.99	1	0.32	0.56	0.83	0.95			0.35	0.45	1	1
	Thorax	HS	0.71	0.85	0.52	0.63	0.97	0.99			0.99	1	0.91	0.97
		MS	0.7	0.77	0.34	0.41	0.83	0.91			0.9	0.95	0.43	0.52
		LS	0.71	0.83	0.01	0.02	0.79	0.91			0.92	0.99	0.49	0.55
Yoganandan et al. 1997	-	Force	0.68	0.89	0.77	0.9	0.74	0.89			0.73	0.88	0.71	0.92

Scatter plot for Ellipse Score and DALW Score for all HBMs for all responses of interest

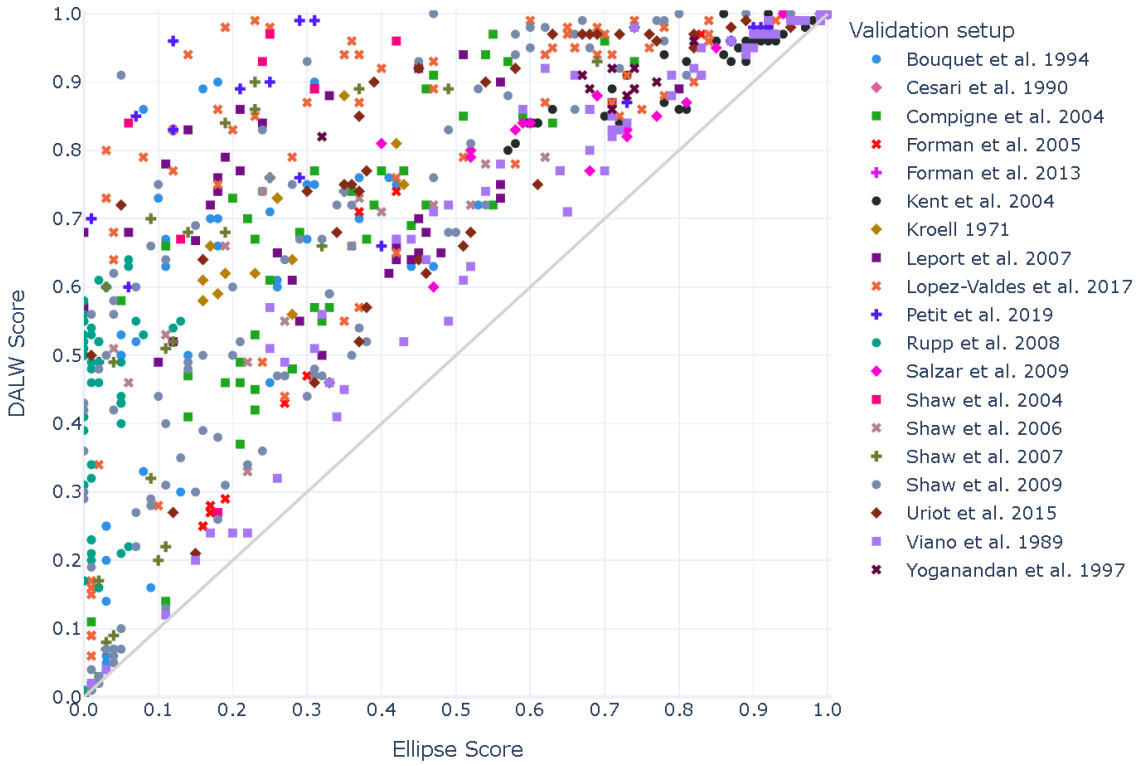

Fig. A 1. Scatter plot illustrating the relationship between the Ellipse Score and DALW Score for all responses of interest of all load cases for all submitted results.

Parameter variation for number of warping control points

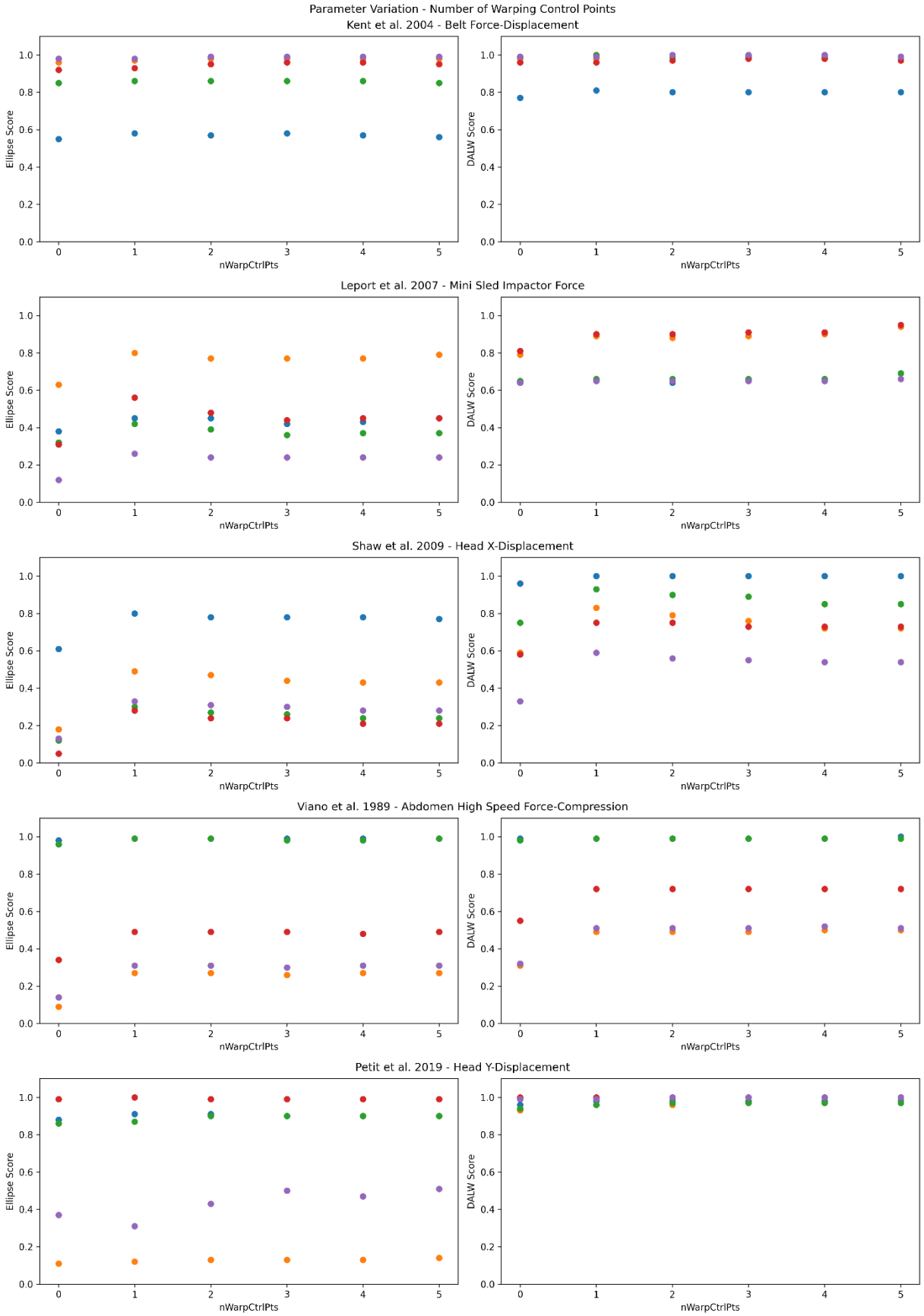

Fig. A 2. Scatter plots of selected responses from five responses of interest showing Ellipse Score and DALW Score for different number of warping control points (*nWarpCtrlPts*). The different colours indicate the different HBMs.

Overview of all HBM-neutral validation setups, load cases, and responses of interest

Table A IV below lists all HBM-neutral validation setups available on OpenVT at the time of writing this study, including the load cases and responses of interest with their evaluation specifics. Note that for rows highlighted in grey, the data available did not allow for the calculation of scores due to insufficient data.

Please refer to the latest version of Euro NCAP's Technical Bulletin CP 550 for the list currently applicable to the respective HBM qualification procedure.

TABLE A IV

OVERVIEW OF ALL HBM-NEUTRAL LOAD CASES AVAILABLE ON OPENVT AND THEIR EVALUATION SPECIFICS

Configuration	Evaluation	x_min	x_max	Min. Ellipse Score	Min. DALW Score
Bouquet et al., 1994,v0.3.1					
Pelvis Low Speed	Force – Time	0	36 ms (end of test data)	-	-
Pelvis High Speed	Force – Time	0	36 ms (end of test data)	0.14	0.46
Sternum Low Speed	Force – Time	0	57 ms (end of test data)	-	0.50
Sternum High Speed	Force – Time	0	57 ms (end of test data)	-	0.66
Cesari & Bouquet, 1990, v0.1.5					
Insufficient test data available	-	-	-	-	-
Compigne et al., 2004; v0.2.7					
3 m/s 0°	*Insufficient test data available*	-	-	-	-
4 m/s 0°	Force – Deflection	0	Cut unloading at 80% of max deflection	-	0.11
6 m/s 0°	*Insufficient test data available*	-	-	-	-
1.5 m/s 0°	Force – Deflection	0	Cut unloading at 80% of max deflection	0.11	0.14
1.5 m/s 15°	Force – Deflection	0	Cut unloading at 80% of max deflection	-	-
1.5 m/s -15°	Force – Deflection	0	Cut unloading at 80% of max deflection	-	-
Forman et al., 2005; v0.2.4					
Tabletop Test	Force-Chest Compression	Cut data so that only data ≥ 5% of f_max remains	Up to max. arc length of characteristic average	0.16	0.25
Forman et al., 2013; v0.7.6					
Insufficient test data available	-	0	250 ms	-	-
Kent et al., 2004 – v0.8.9					
Hub	Force - Compression	Cut data so that only data ≥ 5% of f_max remains	max. arc length of characteristic average	0.70	0.85
Belt	Force - Compression	Cut data so that only data ≥ 5%	max. arc length of characteristic	0.57	0.80

		of f_max remains	average		
Double-Belt	Force - Compression	Cut data so that only data ≥ 5% of f_max remains	max. arc length of characteristic average	0.61	0.84
Distributed Belt	Force - Compression	Cut data so that only data ≥ 5% of f_max remains	max. arc length of characteristic average	0.78	0.86
Kroell et al., 1971 (taking Lebarbé & Petit (2012) into account); v0.2.5					
Hub Impact High Speed	Force-(Hub) Displacement	0	Cut unloading at 80% of max deflection	0.16	0.58
Leport et al., 2007; v0.2.3					
Hub Test	Impactor Force – Time	0	44 ms (end of test data)	-	0.55
	PSIS Force – Time	0	44 ms (end of test data)	-	0.49
Mini Sled	Impactor Force – Time	0	44 ms (end of test data)	0.26	0.50
	PSIS Force – Time	0	44 ms (end of test data)	0.11	0.52
Lopez-Valdes, 2017; v0.6.5					
Restraint Condition A	Head X Displ-Time	0	159 ms (end of test curves)	-	0.64
	T8 X-Displ – Time	0	159 ms (end of test curves)	-	0.68
	T1 X-Displ – Time	0	159 ms (end of test curves)	0.23	0.85
	Pelvis X-Displ Time	0	159 ms (end of test curves)	0.62	0.94
Restraint Condition B	Head X Displ-Time	0	159 ms (end of test curves)	0.58	0.79
	T8 X-Displ – Time	0	159 ms (end of test curves)	-	-
	T1 X-Displ – Time	0	159 ms (end of test curves)	-	-
	Pelvis X-Displ Time	0	159 ms (end of test curves)	-	0.57
Petit et al., 2019; v0.7.9					
Farside Sledtest	Head y displacement – time	0	200 ms (end of simulation)	0.12	0.87
	Head z displacement - time	0	200 ms (end of simulation)	-	0.49
Rupp et al., 2008; v0.1.14					
1.2 m/s	Force – Time for left and right knee	0	69 ms (end of loading)	-	-
3.5 m/s	Force – Time for left and right knee	0	50 ms (end of loading)	-	0.20
4.9 m/s	Force – Time for	0	40 ms (end of	-	0.16

	left and right knee		loading)		
Salzar et al., 2009; v0.5.3					
0.5 m/s	Force - Compression	0	max. arc length of characteristic average	0.40	0.60
1 m/s	Force - Compression	0	max. arc length of characteristic average	0.68	0.77
G. Shaw et al., 2004; v0.3.2					
Up to 50% chest depth	*Insufficient test data available*	-	-	-	-
Up to 30% chest depth	Force – Compression	Force >0	Unloading cut at max. force	-	0.27
J. M. Shaw et al., 2006; v0.4.1					
Lateral impact (90°)	Force- Compression	0	Displacement and force at 100 ms	-	0.49
Oblique impact (60°)	Force - Compression	0	Displacement and force at 100 ms	-	0.33
G. Shaw et al., 2007; v0.4.0					
Non-injurious U – 3rd left rib	Force Displacement	0	Up to max. arc length of characteristic average	-	-
Non-injurious M – mid sternum	Force Displacement	0	Up to max. arc length of characteristic average	-	-
Non-injurious L – 6th right rib	Force Displacement	0	Up to max. arc length of characteristic average	-	0.49
Injurious	*Insufficient test data available*	-	-	-	-
G. Shaw et al., 2009 – Gold Standard Test; v0.9.0					
Configuration 1	*No trajectories available*	-	-	-	-
Configuration 2	Head x displacement – time	0	240 ms (end of test curves)	0.24	0.59
	Head y displacement – time	0	240 ms (end of test curves)	-	0.43
	Head z displacement – time	0	240 ms (end of test curves)	-	0.19
	T1 x displacement – time	0	240 ms (end of test curves)	-	0.27
	T1 y displacement – time	0	240 ms (end of test curves)	-	0.22
	T1 z	0	240 ms (end of test	-	-

	displacement – time		curves)		
	T8 x displacement – time	0	240 ms (end of test curves)	-	0.40
	T8 y displacement – time	0	240 ms (end of test curves)	-	0.34
	T8 z displacement – time	0	240 ms (end of test curves)	-	-
	Pelvis x displacement – time	0	240 ms (end of test curves)	-	0.29
	Pelvis y displacement – time	0	240 ms (end of test curves)	-	0.36
	Pelvis z displacement – time	0	240 ms (end of test curves)	-	-
	Chest Deflection - Time	0	150 ms (end of test curves)	-	0.52
Uriot et al. 2015; v0.6.7					
Seat Configuration Front	Hip x-displ vs. time	0	110 ms (max. test data)	0.67	0.95
	Pelvis rotation vs. time			-	0.21
Seat Configuration rear	Hip x-displ vs. time	0	110 ms (max. test data)	-	0.72
	Pelvis rotation vs. time			0.12	0.427
Viano, 1989; v0.6.7					
Hip High Speed	Force-Compression	0	max. force	0.25	0.51
Hip Medium Speed	*Insufficient test data available*	-	-	-	-
Hip Low Speed	Force-Compression	0	max. force	0.25	0.57
Abdomen High Speed	Force-Compression	0	max. force	0.20	0.24
Abdomen Medium Speed	Force-Compression	0	max. force	-	-
Abdomen Low Speed	Force-Compression	0	max. force	0.32	0.45
Thorax High Speed	Force-Compression	0	max. force	0.51	0.61
Thorax Medium Speed	Force-Compression	0	max. force	0.22	0.24
Thorax Low Speed	Force-Compression	0	max. force	-	-
Yoganandan et al., 1997; v0.5.7					
Hub Impact	Force-Compression	0	Unloading cut at 80% of max. displacement	0.32	0.82