Development of New Objective Indicators to Identify Submarining Risk in Simulations with ATDs and HBMs

Marius Rees, Ryusuke Asahi, Philipp Wernicke, Steffen Peldschus

Abstract This study introduces two objective, automated useable indicators to virtually assess submarining risk, addressing limitations of visual analysis and the Euro NCAP force drop criterion. While prior work has explored submarining mechanisms in physical tests, universally applicable virtual detection methods remain lacking. A parameter study including 800 simulations using Anthropomorphic Test Devices (ATDs) and Human Body Models (HBMs) across generic seating environments were analysed. Among several candidates based on distance, force, or angle metrics, two indicators emerged as most reliable: Nearest Node Dynamic (NND) and Nearest Node Pretensioning (NNP). Both quantify the distance between lap-belt nodes and the anterior superior iliac spine (ASIS), with characteristic value changes indicating submarining. Indicators were evaluated in both global and local coordinate systems, the latter defined by ASIS and anterior inferior iliac spine (AIIS). Contrary to expectations, the local system offered no clear performance advantage. NND and NNP accurately distinguished submarining from non-submarining cases, with few false classifications. These indicators offer scalable, robust tools for virtual safety assessments in varied postures, seat configurations, and population studies.

Keywords Submarining injuries, human body models (HBMs), anthropometric test devices (ATDs), autonomous driving, lap belt

I. INTRODUCTION

The introduction of autonomous vehicles is expected to reshape automotive design, enabling new seating configurations focused on passenger comfort. Reclined postures in particular raise safety concerns due to the increased risk of submarining - a phenomenon where the lap belt slides over the iliac crest, potentially causing internal injuries during a crash [1-5]. This risk increases in reclined positions, because the altered posture rotates the pelvis into a more horizontal orientation, reducing lap-belt contact with the iliac crest. As a result, the belt's ability to restrain the lower torso effectively is compromised [2][6-7]. Simulation studies confirm this elevated risk and highlight contributing factors such as pelvic soft tissue and restraint configurations [8-11]. Most of these studies detect submarining through qualitative visual methods, e.g. result image analysis or tracking belt and ASIS marker trajectories. Although these methods yield valuable insights, they lack scalability, objectivity, and automation, limiting their application in large simulation studies. An automated and objective usable, scalar indicator could streamline and standardise submarining evaluation.

Future robustness studies of new (adaptive) restraint systems will increasingly include diverse occupant models varying in sex, size, and shape [12-16], often requiring hundreds of simulations. Traditional evaluation methods, like visual inspection or the Euro NCAP force drop criterion (≥1 kN within 1 ms) [17], often lack the objectivity and automation required for large-scale, simulation-based investigations, as well as the generalisability to be applied to both ATDs and HBMs. The force drop threshold of 1 kN within 1 ms is often not met in HBM simulations even in clear submarining scenarios, due to more gradual force reductions, making the criterion prone to false negatives as discussed later. Efficient, objective indicators are therefore essential for large-scale virtual testing.

Although HBMs are widely used for reclined posture studies, ATDs remain central to regulatory and consumer testing due to their standardisation and repeatability. Submarining can also occur in ATDs, even in upright postures [18-20], reinforcing the need for methods applicable across both model types.

This study aimed to develop and evaluate objective, automated submarining indicators for use in simulations with ATDs and HBMs. Twenty candidate indicators were tested across over 800

M. Rees (e-mail: marius.rees@bmw.de) and P. Wernicke are safety engineers at BMW AG in Munich. R. Asahi is a Technical Leader at Vehicle Development Division, Mazda Motor Corporation, Japan. S. Peldschus is Professor at the LMU University of Munich.

simulations, encompassing various environments, postures, and occupant types. Two indicators demonstrated strong performance and are presented in detail. Their predictive accuracy was validated against manually identified submarining events using video analysis. The goal is to offer a reliable, scalable tool for objective safety assessment in virtual testing scenarios.

II. METHODS

Initially, 20 potential indicator definitions were defined and evaluated in 800 simulations, covering four seating environments, two postures (upright and reclined), two HBM anthropometries, and one ATD (Table I). Two of the environments allowed both postures, resulting in six distinct seating configurations. THUMS v4.1 50th percentile model [21-22] was integrated into all six configurations as the baseline. In a second step, the THUMS v4.1 05th percentile and the THOR AV 50M ATD [23] were each integrated into one of the configurations to examine whether the baseline results could also be observed with other occupant models or anthropometries. A simulation matrix of eight configurations was formed based on these variables. In each, 100 simulations were run with varied input parameters such as friction coefficients, pretensioner ignition times, or sled acceleration levels. The variability aimed to generate both submarining and non-submarining cases, accepting that some scenarios might include unrealistic boundary conditions.

All 20 indicators were applied consistently across the entire simulation matrix. Four indicators emerged as particularly promising in terms of distinguishing between submarining and non-submarining cases. These four indicators demonstrated the most robust performance and were therefore selected for detailed presentation and analysis in this study. A comprehensive overview of all 20 indicators will be presented in a future publication, including results from additional occupant models across all six environments.

TABLE I
SIMULATION MATRIX, WITH NUMBER OF SIMULATIONS (N=100) INVESTIGATED PER EACH CONFIGURATION

Occupant model	Seating environments and seating postures					
	Upright Posture			Reclined Posture		
	LAB/CEESAR semi-rigid Seat	LAB/CEESAR-Integrated NHTSA Seat	Tandem-Sled	LAB/CEESAR semi-rigid Seat	LAB/CEESAR-Integrated NHTSA Seat	Comfort Seat
THUMS v4.1 50th	100	100	100	100	100	100
THUMS v4.1 05th	--	--	--	--	100	--
THOR AV 50M	--	--	--	--	100	--

dark gray = baseline study, light gray = validation study.

Model environments and used occupant models

The six seating configurations are illustrated in Fig. 1. One of the considered environments is the generic semi-rigid seat environment by [24], which was used in several PMHS studies [1-3][25]. The semi-rigid seat is composed of two adjustable articulated plates, each of which can be independently modified in terms of geometry and spring stiffness to replicate a specific production seat. In the following it will be referred to as “LAB/CEESAR semi-rigid seat” [A1, A2].

Initial ATD and PMHS tests using the semi-rigid seat setup [24-25] involved upright postures with a 22 degrees backrest angle. Later studies on reclined postures used the same setup with backrest angles up to 50 degrees [1][9], applying only the “front seat” configuration from [24-25], which includes seat pan and ramp angles of roughly 15 and 30 degrees relative to the horizontal, respectively. For our reclined simulations, we adopted the finite element environment from [9] without changes. For upright simulations, both seat components were rotated counterclockwise to approximate a more vertical seating position, resulting in reduced angle values of about 7 and 18 degrees, respectively, which were not used in [24]. The backrest angle was maintained at 22 degrees rearward of vertical, consistent with [24].

To better represent occupant to belt to seat interaction, we integrated the setup from [24] with a more realistic seat model: a 2019 Honda Odyssey seat from NHTSA [26]. Rigid seat elements were replaced with the cushion and backrest from [26]. Other elements, such as the restraint system and sled pulse, remained unchanged. This setup, with seat angles of 18 degrees (upright) and 45 degrees (reclined), is termed “LAB/CEESAR Integrated NHTSA Seat” ([B1,B2]).

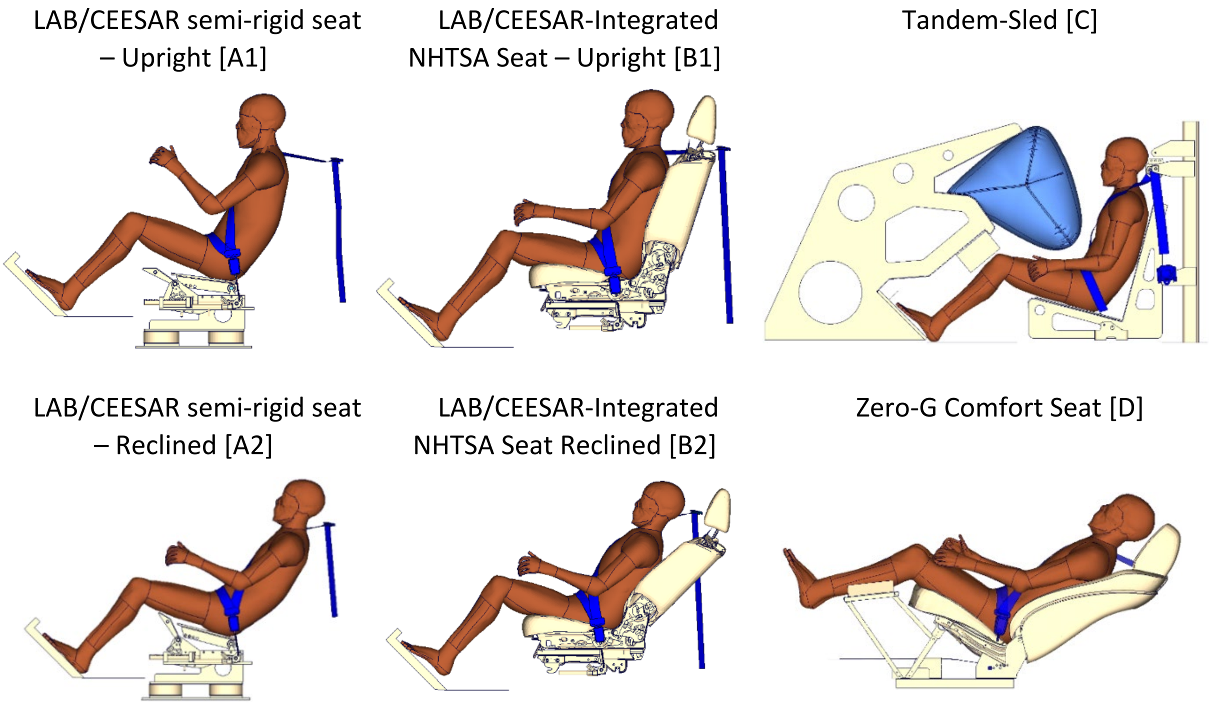

Fig. 1. Overview of all six seat configurations investigated, including a positioned THUMS v4.1 50th. LAB/CEESAR semi-rigid seat [A1,A2] / LAB/CEESAR Integrated NHTSA seat [B1,B2] / Tandem Sled [C] / Zero-G Comfort Seat [D]. *Top:* three upright seating postures. *Bottom:* three reclined postures.

Another environment is the “Tandem” rigid sled for upright seating [27], featuring an adjustable seat plate and production restraint system with pretensioner, load limiter, and pre-inflated airbag. The final environment, the “Zero-G comfort seat”, is a reclined concept prototype by BMW [28], used previously to study lumbar spine loading with HBMs and ATDs.

The study was conducted in two steps: a baseline study and a validation study as highlighted in dark and light gray in Table I. As the focus lies on defining and assessing submarining indicators, detailed descriptions of the occupant models, their positioning and validation, as well as a more comprehensive overview of the seating configurations, are omitted here but can be found in the cited references. The baseline analysis involved applying the 50th percentile THUMS model across all seating environments and postures (Table I), while the follow-up study tested the indicator robustness to other surrogate models or anthropometries, namely the 05th percentile THUMS HBM and the THOR AV 50M ATD, aiming to confirm the generalisability of the baseline findings.

Parameter Study

For each of the introduced configurations, a parameter study was conducted to generate a set of diverse scenarios concerning the interaction between the belt and the occupant. This variation is necessary to assess the feasibility of potential evaluation metrics. Accordingly, 100 simulations were performed per configuration, with input parameters systematically varied based on the Sobol sampling method [29], which ensures a uniform and representative distribution across the given continuous parameter range. The input parameters are summarised in Table II. Adjustments included scaling the acceleration pulse between 75% to 125% of its original magnitude, modifying seat cushion stiffness (only in [B1] and [B2], Fig. 1) and varying the load limiter force as well as the airbag vent orifice coefficient (only in [C], Fig. 1). In the LAB/CEESAR semi-rigid seat configuration ([A1],[A2]), both for upright and reclined postures the belt anchorage positions were shifted 50 mm rearward.

TABLE II

SUMMARY OF INPUT PARAMETERS, THEIR RANGES, AND THEIR USAGE IN DIFFERENT SEAT SET-UPS.

			used in ..			
x	Parameter	Range	[A1,2]	[B1,2]	[C]	[D]
1	Acceleration Scaling [-]	0.75 – 1.25	X	X	X	X
2	Friction Belt to Skin [-]	0.10 – 1.0	X	X	X	X
3	Friction Seat to Skin [-]	0.10 – 1.0	X	X	X	X
4	Activation Time Belt Pretensioner [ms]	5 – 125	X	X		X
5	Stiffness Seat Cushion Scaling [-]	0.5 – 1.5		X		X
6	Seat-belt Load Limiter Scaling [-]	0.5 – 1.5			X	
7	Airbag Vent Orifice Coefficient [-]	0.5 – 4.0			X	
8	Anchorage Position [mm]	+ 50	X			

Pair-plots illustrating the variation across all parameters can be found in Fig. A1 in Appendix A. All simulations were executed with the FE-solver LS-DYNA MPP R12.2 [ANSYS LSTC, Livermore, CA].

Description of the two Submarining Indicators

Both indicators introduced in the following section share a common approach: they measure the distance from the two ASIS reference points to a specific reference point on the lap belt at each time step. Due to the continuous evaluation throughout the entire simulation, the first indicator is also referred to as "Nearest Node *Dynamic*". The analysis is conducted on both the left and right sides of the body.

(a) "Nearest Node Dynamic" (NND)

The NND is acquired in three-steps, repeated at each time step. In the first step, a narrow three-dimensional selection box is defined around the ASIS point, with minimal width in the global y-direction. All lap-belt nodes within this box are identified, forming a point cloud of lap-belt nodes positioned directly "in front of" ASIS (Fig. 2). In the second step, the node within this point cloud that is farthest in the global z-direction (in vertical direction) from the ASIS point is determined (Fig. 2 *left*, with the measurement (2) indicated in yellow and the selected node in green). Finally, in the third step, the distance in the x-direction between the ASIS point and this specific belt node is calculated (Fig. 2 *right*, measurement (3)). In this context, the term "nearest" specifically relates to the location in the y-direction only (see additional comment in Appendix K).

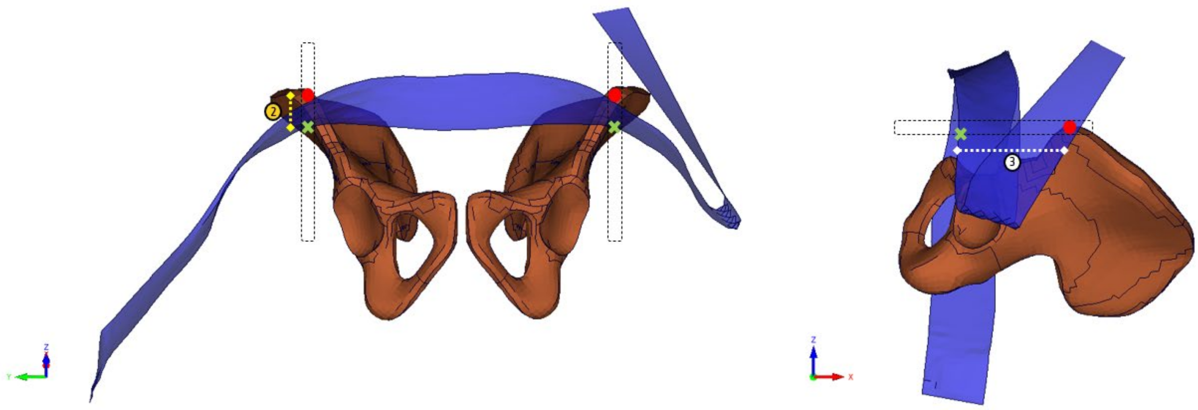

Fig. 2. Definition of first Indicator "Nearest Node Dynamic".

If the measured x-distance changes sign, transitioning from positive to negative, this indicates that the nearest lap-belt node has moved "behind" the ASIS, suggesting that submarining might have occurred. These steps are repeated at each output interval of 1 ms throughout the entire simulation duration. The x-distance measured at every output state is then plotted as a function of time. Since the selection of lap-belt nodes may change at each time step, the resulting plot exhibits a step function characteristic. The minimum of this curve is then extracted to determine the NND value.

Two output plots are generated. The first is a 2D time series showing the measured x-distance over time (Fig. 3).

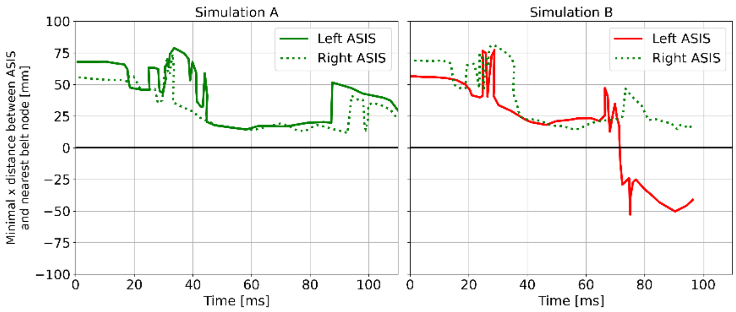

Fig. 3. X-distance between ASIS and nearest belt node (NND) for two simulations. *Left*: Simulation A, no submarining. *Right*: Simulation B, submarining on left side (red line).

The second is a scatter plot displaying the minimum x-distance from each simulation (Fig. 4). Red markers denote submarining cases confirmed via video analysis, green markers indicate non-submarining cases. The scatter plot is especially useful for large-scale studies, offering a quick overview across many simulations, while the time series supports detailed analysis of individual runs.

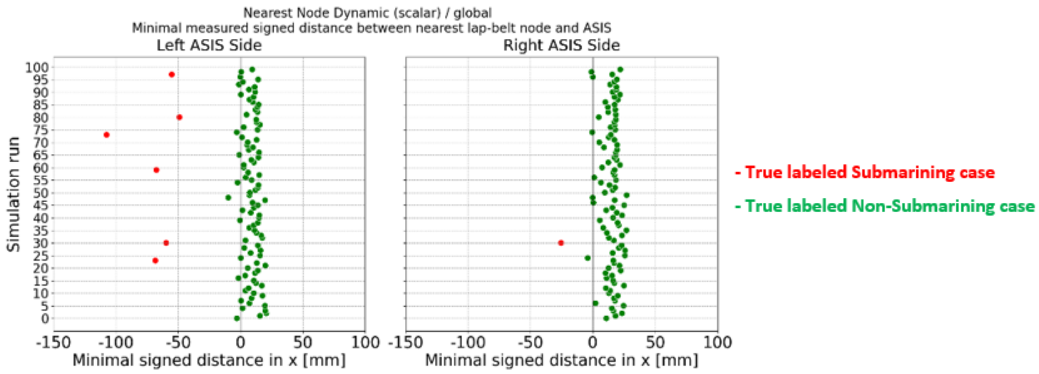

Fig. 4. Minimum measured distances over 100 simulation runs based on values from Fig. 3. As an example, Simulation 80 corresponds to the run shown as Simulation B in Fig. 3.

(b) "Nearest Node Pretensioning" (NNP)

In contrast to the previously discussed dynamic implementation, an alternative approach does not involve scanning for a point cloud of belt nodes at each output state. Instead, it focuses on the ignition time of the pretensioner. Shortly after this time, a 3D box defined as in the NND method is used to identify the belt node with the smallest x-coordinate, corresponding to the farthest belt node from the ASIS (see coordinate system sketched in Fig. 2). This specific belt node is then fixed and tracked for the remainder of the simulation. At each timestep, the minimum x-coordinate difference between this single, preselected belt node and the ASIS is evaluated. As before, a change in sign in this distance may indicate a potential submarining event. Same plots as described for the first indicator are generated.

(c) Local Coordinate System

Both indicators were not only evaluated within the global coordinate system but also in a local anatomical coordinate system to account for pelvis orientation in reclined postures, which can distort results using global axes. This "ASIS–AIIS" system is defined using the Anterior Superior Iliac Spine (ASIS) and Anterior Inferior Iliac Spine (AIIS) landmarks on both body sides (Fig. 5). A plane is defined through these points, with its center forming the origin. The z-axis is normal to the plane, the y-axis connects the ASIS points, and the x-axis is orthogonal to both. The calculation for indicator (a) and (b) remain the same as in the global system, but with the coordinate axes adjusted accordingly at each time step.

Evaluation timeframes vary by simulation and were conducted here up to shortly after rebound. In terms of instrumentation, ASIS, AIIS and lap-belt nodes must be monitored, with distances and coordinate transformations computed using a Python script. These scripts have already been integrated into a user-friendly GUI, which will be freely accessible within the TUC [30].

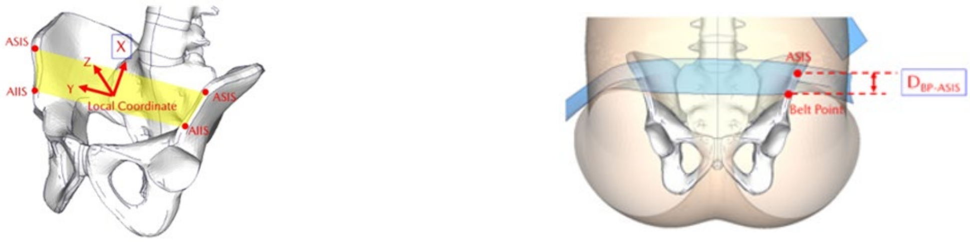

Fig. 5. Definition of the local coordinate system “ASIS-AIIS”.

Post-Processing – Evaluation of the Parameter Study, Video Analysis and Rating of the Indicators

To evaluate indicator performance, each of the 100 simulations per configuration were visually analysed in three perspectives (left, right, top view) to determine whether submarining occurred or not. The results provided the true labels used for the red/green classification shown in Fig. 3 and Fig. *4*.

To evaluate the indicators both statistically and practically, a two-step approach was applied. First, the Wilcoxon rank-sum test [31-32] was used to assess whether the distributions of submarining and non-submarining cases differ significantly. Based on this, the Matthews Correlation Coefficient (MCC) was then calculated to quantify classification performance [33-35].

The Wilcoxon rank-sum test was applied to assess whether the submarining indicator values show a statistically significant difference between submarining and non-submarining cases, thereby establishing a statistical basis for differentiation. This non-parametric statistical test compares the distributions of two independent groups (submarining vs. non-submarining) by calculating a p-value, which quantifies the probability that the observed differences occurred by chance under the null hypothesis of no difference. The test does not assume normal distributions, making it suitable for our data. To improve interpretability, the p-values were plotted in logarithmic scale ($-\log_{10}(p)$) to emphasise small p-values (i.e. highly significant results), making it easier to visually distinguish strong vs. weak statistical significance. For example, a p-value of 0.001 becomes 3 on this scale (highly significant), while a non-significant p-value of 0.5 is below 0.3 (non-significant).

To complement the statistical test, the MCC was calculated to evaluate each indicator’s classification performance across different environments. Binary predictions were generated using a threshold of 0 mm, and confusion matrices were constructed to compare predicted and actual outcomes [36]. MCC is a balanced performance metric that combines all four entries of the confusion matrix - true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) - making it suitable even for imbalanced datasets such as ours. It is defined as:

$$\mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

The MCC ranges from -1 (complete disagreement) to +1 (perfect prediction), with 0 indicating random classification. This enables a robust assessment of how well an indicator distinguishes submarining from non-submarining cases.

III. RESULTS

First, the statistical significance of the indicator values was assessed using a Wilcoxon rank-sum test. The results are shown in Fig. D1 in Appendix D. A low p-value (corresponding to a high $-\log_{10}(p)$ value in Fig. D1) indicates that the indicator can statistically differentiate between submarining and non-submarining cases. The consistently low p-values on the left ASIS side especially for the last four environments [C], [A2], [B2] and [D] suggest that the indicators reliably separate submarining from non-submarining cases on that side. On the right side, p-values are not significant (near 0 on log scale) especially in the first two upright environments [A1] and [B1]. These environments included only one or two submarining cases on the right side, which limits the statistical contrast between submarining and non-submarining groups. These findings support continuing with the Matthews Correlation Coefficient (MCC) to further evaluate the indicators classification performance.

Baseline study

Simulation snapshots for submarining cases in environments [C] and [B2] are provided in Appendix B and C (Fig. B1, Fig. C1). Due to space limitations, the following results focus on configuration [B2] from the baseline study, illustrating the performance of NND and NNP indicators with THUMS v4.1 50th. Summaries for all six environments are shown in Appendix E (Fig. E1 to Fig. E6). Table III lists the distributions from the video analysis for submarining and non-submarining cases.

TABLE III

SUMMARY OF SUBMARINING AND NON-SUBMARINING CASES FOR SIX CONFIGURATIONS

	Load case		Submarining Cases	Non-Submarining Cases
THUMS v4.1 50th	[A1]	LAB/CEESAR semi-rigid seat (Upr.)	6	94
	[B1]	LAB/CEESAR-Integrated NHTSA Seat	4	96
	[C]	Tandem Sled	19	81
	[A2]	LAB/CEESAR semi-rigid seat (Recl.)	22	78
	[B2]	LAB/CEESAR-Integrated NHTSA Seat	19	81
	[D]	Comfort Seat	58	42

The analysis of the reclined seating posture [B2] includes 19 submarining and 81 non-submarining cases, with indicator predictions summarised in Fig. 6. For all four evaluated indicators a separation between both scenarios can be visually observed, although the degree of separation varies for each indicator.

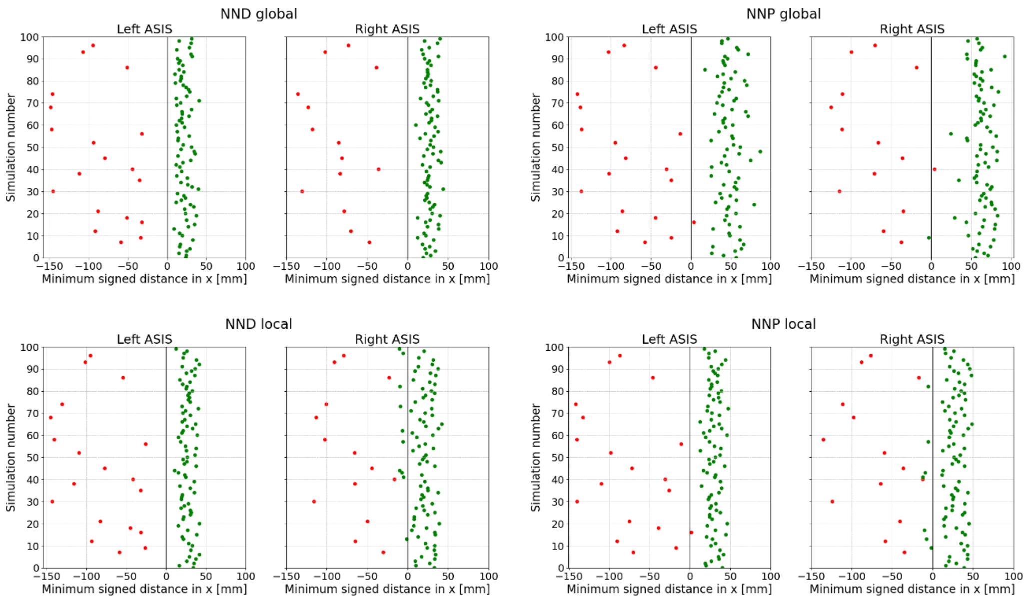

Fig. 6. Comparison of indicator predictions to true label information for configuration [B2]. NND global (*top left*), local (*bottom left*), NNP global (*top right*), NNP local (*bottom right*).

The global NND clearly distinguishes submarining cases, which all have negative values meaning that a negative signed distance was calculated for these cases. The numerous non-submarining cases are mostly clustered near the zero line, with some slightly exceeding it especially for the local evaluated NND. Ten non-submarining cases were misclassified as submarining, occurring only on the right ASIS side (e.g. run 41 or 99). A similar pattern is observed in both NNP evaluations, where values shift slightly more to positive. Notably, runs 16 (left ASIS) and 40 (right ASIS) show values near zero in NNP global, while the NND values for these runs are nearly 50 mm more negative, at -39 mm and -49 mm respectively. Overall, only the global NND plot shows a clear separation between submarining and non-submarining cases. Similar findings for the indicators were also observed in the other five environments (Fig. E1 to Fig. E6).

The clear separation between the two cases is also reflected in the confusion matrix in Table IV, shown here again only for environment [B2]. The strong predictive performance of the global NND is evident in its mean MCC of 1.0, while the global NNP is lower at 0.94. Both local evaluations yield lower values with 0.86 and 0.88 due to a higher number of false positives. Confusion matrices for the other five environments can be found in Tables G.I to G.V in Appendix G.

TABLE IV
CONFUSION MATRIX FOR ENVIRONMENT [B2] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT

Indicators		TP	TN	FP	FN	MCC	Ø MCC*
NND global	Left	19	81	0	0	1	1
	Right	14	86	0	0	1	
NND local	Left	19	81	0	0	1	0.86
	Right	14	76	10	0	0.718	
NNP global	Left	18	81	0	1	0.967	0.94
	Right	13	85	1	1	0.917	
NNP local	Left	18	81	0	1	0.967	0.87
	Right	14	79	7	0	0.783	

* The value in the last column represents the mean MCC derived from the individual MCC values of the left and right ASIS.

Based on these tables, an average MCC across all environments was calculated for each indicator. These results are summarised in Table V and visualised in Fig. 7. Bar heights represent the average of the left and right ASIS MCC per environment, while the dashed line indicates the overall mean across all six environments.

TABLE V
MC-COEFFICIENTS FOR ALL FOUR INDICATORS OVER ALL SIX CONFIGURATIONS

Indicator	MCC for environment ..						
	[A1]	[B1]	[C]	[A2]	[B2]	[D]	Ø MCC
NND global	1	1	1	1	1	0.94	0.99
NND local	1	1	0.98	1	0.86	0.92	0.96
NNP global	0.78	0.67	0.96	0.97	0.94	0.91	0.87
NNP local	0.51	0.69	0.98	0.97	0.87	0.91	0.82

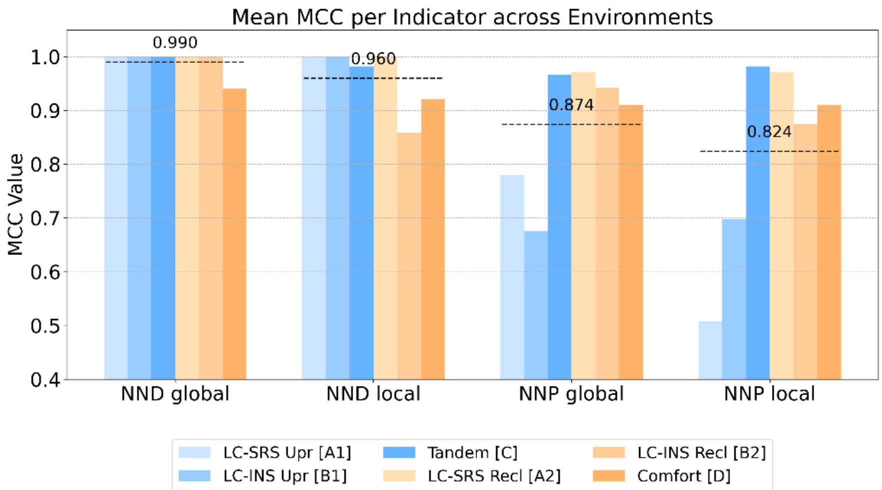

Fig. 7. MCC of all four indicators across six setups (average of l/r ASIS); dashed line = overall mean

The global NND shows the highest predictive performance with an MCC of 0.99, followed by its local version with an average of 0.96. The NNP indicator shows slightly lower mean values below 0.88. Especially in the first two upright postures, where NND shows correlations of 1.0, the NNP instead shows lower correlations of 0.78 and 0.66. In the reclined postures, both global indicators follow similar patterns, although NND yields higher MCC overall. Switching to the local coordinate system in upright postures does not affect NND but worsens or remains nearly unchanged for NNP. In the reclined positions, MCC for both indicators either decrease slightly or remain stable when switching to the local coordinate system.

Validation study

To validate the baseline findings, the LAB/CEESAR-Integrated NHTSA seat [B2] in a reclined configuration was used, placing the THOR AV 50M ATD alongside the 05th percentile HBM. Submarining and non-submarining results are summarised in Table VI. Fig. F1 and Fig. F2 in Appendix F display the scatter plots for both models, the MCC values are provided in Fig. J1 in Appendix J.

TABLE VI

SUMMARY OF SUBMARINING AND NON-SUBMARINING CASES IN VALIDATION STUDY

	Environment	Submarining	Non-Submarining
THUMS v4.1 05th	[B2] LAB/CEESAR-Integrated NHTSA Seat	20	80
THOR AV 50M	[B2] LAB/CEESAR-Integrated NHTSA Seat	50	50

In both cases, the results are consistent with previous findings, as summarised in Table VII. For the AF05 THUMS (see also Fig. F1), the NND shows clear separation with MCC of 1.0 for both sides. The NNP performs well on the right, but two false positives on the left reduce its MCC slightly. In the local system, values remain largely unchanged. For the THOR AV 50M (Fig. F2), the global NND achieves the highest MCC (0.96). Both local indicators and the global NNP show less separation compared to NND, with MCC decreasing slightly or staying the same.

TABLE VII

MC-COEFFICIENTS FOR ALL FOUR INDICATORS OVER BOTH VALIDATION STUDIES

	THUMS v4.1 AF05	THOR AV 50M
NND global	1	0.96
NND local	1	0.94
NNP global	0.97	0.92
NNP local	0.95	0.92

Besides MCC, other classification metrics like Recall, Prediction and F1-score were also calculated for all results, included in Appendix I and J and referenced in the following discussion chapter.

IV. DISCUSSION

First, the statistical evaluation methods and their limitations are discussed, followed by points addressing the indicator evaluation methodology and its specific limitations.

The statistical assessment of the proposed indicators began with the Wilcoxon rank-sum test to determine if indicator values differed significantly between submarining and non-submarining cases. This non-parametric test is suitable for comparing two independent groups without assuming normality and is robust to outliers, making it appropriate for simulation-based biomechanical data that can be skewed or contain extremes. It also tolerates unequal sample sizes, relevant here due to the imbalance between submarining and non-submarining cases across six environments (Table III).

Results showed consistently low p-values (high $-\log_{10}(p)$) on the left ASIS side in most environments, confirming the indicators' ability to distinguish the groups there. On the right side, the p-values remained high in the first two upright environments [A1] and [B1], reflecting the absence of statistically significant differences. This limitation is attributed to the very small number of submarining cases in those settings, which reduces the statistical strength of the test and makes it harder to detect meaningful differences. In such cases, the absence of significance should not be misinterpreted as a failure of the indicator but rather as a data limitation.

While the Wilcoxon test is simple and robust, it only ranks values and does not reflect the magnitude of differences. It also does not indicate how well a given threshold or model performs in practice when used to classify new data. Therefore, to quantify the actual classification performance of each indicator, the MCC was applied in a subsequent step. Second, in environments with very low case numbers (e.g. only four or six submarining cases), the test strength is limited. Future studies could validate the findings through alternative resampling methods such as permutation tests, which can provide additional robustness in small-sample, imbalanced settings [37]. Despite these limitations, the use of the Wilcoxon test in this context was an appropriate and statistically valid choice. It provided an initial assessment of group separability and justified the subsequent application of classification metrics like the MCC.

After confirming statistical separation with the Wilcoxon test, the indicators were further evaluated using MCC. MCC considers all confusion matrix values (TP, TN, FP, FN) and is useful for imbalanced datasets such as the one in this study. While all four indicators showed satisfactory predictive accuracy across six configurations (Fig. 7), MCC scores varied between them. Generally, NND had a higher average MCC than NNP (Table V). In the three upright seats, NNP showed more variation, especially in positive values (Fig. E1–E3). This is because NNP uses only one fixed lap-belt node, specifically the node nearest to ASIS at pretensioner ignition throughout the full simulation time. For example, simulation A measures a node at 2 ms, while simulation B uses one at 95 ms, causing greater scatter in NNP results. In contrast, NND always uses the nearest node, resulting in less variability.

A further limitation arises if the selected pretensioner node is behind the ASIS initially, causing negative x-coordinates and early negative values in the NNP curve. Since the classification threshold is zero, this may falsely trigger submarining predictions. This effect is stronger in the local coordinate system, where the origin is at the ASIS-AIIS plane, increasing misleading early negatives. These issues contribute to the lower MCC scores observed for NNP compared to the more robust NND indicator.

The reliability of the NNP indicator strongly depends on the timing of the pretensioner activation. In cases with delayed or no pretensioner activation, the belt may not be sufficiently coupled to the pelvis which can affect the interpretability of the NNP measure. It is therefore beneficial to apply NNP in scenarios with early belt engagement. Future vehicle concepts, especially autonomous ones with reclined or rotated seating, may involve delayed activations of pretensioner, then making NND the more robust and generally applicable choice for submarining assessment.

The globally evaluated NND indicator shows the highest MCC across all six environments and is therefore the most suitable of the four tested indicators for assessing submarining. It achieves an averaged MCC nearly 0.1 higher than the NNP indicator (0.99 vs. 0.87) and shows better results in every individual environment, both in upright and reclined seating postures. This advantage results from selecting a new seat-belt node at each time step, which better reflects the dynamic belt to pelvis interaction than the fixed-node approach used in the NNP evaluation, an approach that also causes the previously discussed limitations. This is evident not only in Fig. 7 but also in the confusion matrices of all six environments (Tables VIII-XII in Appendix G), where false negatives and false positives occur in both NND and NNP results, though they are more frequent in NNP. False negatives are particularly critical, as they misclassify actual submarining cases as non-submarining and should be avoided. No false negatives occurred with the global NND indicator in any environment, further underscoring its robustness and predictive value. Similar trends were observed in the validation study using the 05th percentile THUMS HBM and the THOR AV 50M ATD. As before, no false negatives occurred for NND global, while a small number appeared for NND local, still fewer than in either NNP result. This aligns with the higher MCC values observed for NND (Table H.I, H.II in Appendix H, Fig. F1, Fig. F2), supporting the applicability of the indicators to other HBM percentiles or occupant surrogates.

While the presented statistical evaluation provides valuable insights, it also comes with certain limitations. The classification was mainly assessed using MCC, which doesn't differentiate between false positives and false negatives. These error types are particularly relevant in safety-critical applications like submarining detection: false negatives (missed submarining cases) may pose serious safety risks, while false positives (incorrectly flagged cases) could trigger unnecessary interventions and reduce user trust. To gain a more nuanced understanding of classification performance, additional metrics such as precision, recall, and F1-score were computed (Fig. I1, Fig. I2 and Fig. I3 in Appendix I for baseline, Fig. J3, Fig. J4 and Fig. J5 in Appendix J for validation study). These quantify the model's ability to detect all relevant cases (recall) and to avoid false alarms (precision), with the F1-score capturing the trade-off between both. All metrics confirmed the same indicator ranking, with NND showing the strongest performance throughout.

It should be noted that all reported classification metrics were calculated based on a fixed threshold of 0 mm. If this decision boundary were to be adjusted, the resulting metric values could improve or worsen accordingly. Another limitation is the use of a fixed threshold (0 mm) for classification, which was chosen heuristically but not optimised. Future work should consider threshold optimisation, e.g. by systematically varying the decision boundary to maximise metrics like MCC or F1-score. An initial approach has already been explored and is documented in Fig. I4 in Appendix I.

Binary metrics such as MCC do not account for prediction uncertainty or gradual risk levels. Probabilistic models or continuous approaches could reflect such nuances more accurately. One methodological approach that could be valuable in future work is the use of Injury Risk Curves (IRC). While not applied in this study, the results from the NND indicator could serve as a basis for such risk modeling. Developing reliable IRCs requires larger datasets, extensive validation and careful handling of influencing factors such as occupant variability and boundary conditions. Given these requirements, IRCs were not pursued here but remain a promising option for future investigations.

Having discussed the statistical evaluation and its limitations, the focus now shifts to the indicator methodology itself. It became apparent that both indicators show comparable responses depending on the coordinate system used for calculations.

On average, MCC values are slightly lower for the local than for the global evaluation (Fig. 7). No consistent pattern emerges between upright and reclined postures: for NND, upright results are similar in both systems, while in reclined postures, particularly in environment [B2], the local evaluation performs slightly worse. For NNP, upright results vary, and in reclined setups, performance remains similar or marginally lower. Although the local coordinate system was expected to improve predictions by moving with the occupant, the observed differences are minor, with no clear advantage. This suggests that the global coordinate system is sufficient and may even provide better predictions, as it remains fixed relative to the vehicle and aligns with the main direction of the seatbelt force or vehicle axis. In contrast, the local coordinate system, defined by the ASIS-AIIS landmarks, moves and rotates with the occupant's pelvis as posture changes. This rotation can absorb part of the belt movement during coordinate transformation, effectively dampening measured displacements. Since the global system captures all translational and rotational movements relative to the vehicle and aligns closely with belt force and occupant movement directions, it may better reflects critical kinematic changes. As a result, calculations using the local coordinate system do not improve prediction accuracy compared to the global system.

A key issue is the selection of ASIS and AIIS points, as no consistent anatomical reference exists. This challenges both HBMs and ATDs and influences indicator evaluation. Pelvic shape and ASIS/AIIS positions vary significantly between individuals [38-40]. Although two percentiles (50th and 05th) were considered, the current indicators may not fully capture this variability, which is important for different population groups beyond average or extreme body types. Future use should account for this, and it would be valuable to study how small changes in ASIS/AIIS definitions affect indicator results.

An additional factor to consider is the time period chosen for the evaluation. The results here were calculated over a long time period, from 0 ms to 90 ms shortly after the rebound phase starts, which explains the relatively high distance values, especially in the negative range. A time window that is too short could distort the results more than a longer time window. The effect of the time period should be explored further in future studies.

The tested environments and the chosen parameter ranges represent another limitation. The number of submarining cases depends on the variation and combination of input parameters. In this study, non-submarining dominated in five of six environments. A more targeted parameter space, for example by including lap belt angle or anchorage position, could increase submarining cases. The lap belt angle is known to influence submarining risk and should be included in future setups, especially with real seats [41-42].

At the beginning of the study, the zero crossing in the distance curve was defined as the threshold for submarining (Fig. 3). This choice seemed reasonable at first and allowed a basic classification between submarining and non-submarining cases. It became evident that this threshold lies within a critical range where many borderline cases occur, situations in which submarining has not yet fully developed but is imminent and thus requires closer inspection by a safety engineer.

This leads to two conclusions. Future studies should include a systematic threshold optimisation, as previously discussed (Fig. I4). Second, the current binary classification may be insufficient for design purposes. From an engineering perspective, especially when developing new restraint systems, a continuous measure indicating proximity to submarining would be more informative. To address this, a more continuous version of the indicator is currently under development. At each time step, the vertical distance to the belt node farthest in the z-direction is evaluated and processed relative to the initial belt width, resulting in a percentage that reflects the remaining belt coverage. For example, a value of 100% means no risk, while 10% indicates high risk because the ASIS is covered by only 10% of the belt width. This continuous output is intended to complement the binary classification presented in this study. An initial version was tested in an internal BMW study for a future reclined seat concept and showed promising results. It provides a simple way to estimate submarining risk more continuously, without the complexity of a full probabilistic risk curve. By combining both output types of the NND/NNP indicator, a compromise between binary classification and continuous assessment is achieved, maintaining interpretability while increasing flexibility. An example output is given in Appendix K (Fig. K1, Fig. K2) and will be further detailed in a future publication by the author.

Another point of discussion is the force drop criterion, which has long been used to identify submarining by detecting a rapid decrease in lap-belt force at the iliac crest [17]. When applied in simulation studies, especially with HBMs, several limitations emerge. In this study's simulations (Appendix L), even clear submarining cases often did not show a sharp enough drop to meet the common threshold (1 kN within 1 ms). Instead, the force typically decreases more gradually, over 5 ms to 10 ms and with smaller magnitudes. This is particularly relevant for HBMs, where soft tissue and belt interaction are modeled in more detail. For example, in Figures Fig. L1-Fig. L3 (Appendix L), the left ASIS force drops by 0.5 kN within 15 ms and 0.35 kN within 5 ms. As a result, submarining may go undetected when relying solely on the force drop criterion, increasing the risk of false negatives. This is especially true when submarining develops slowly, such as under low belt-to-skin friction, where the belt gradually slides over the pelvis without a sudden force drop (see Fig. L1). Another limitation is the assumption of symmetrical behavior, i.e. similar force reduction on both pelvic sides. Simulations show that submarining can also occur asymmetrically, with the belt slipping on one side while force remains on the other (see Fig. L2), making a global threshold insufficient and further increasing the risk of missed cases.

The criterion is reactive, meaning it identifies submarining only after it has started, capturing the mechanical consequence of submarining after it was initiated. In contrast, the new indicators presented in this study are geometry-based, showing geometric relationships between the lap belt and key pelvic landmarks. These indicators can detect critical movements of the belt relative to the pelvis before a full slip occurs, offering a more anticipatory and robust approach. Overall, while the force drop criterion remains a useful reference, it shows limitations. The proposed indicators provide a more reliable and flexible tool for detecting submarining risk, especially in diverse seating conditions and across different occupant models.

V. CONCLUSION

The predictive accuracy of two objective submarining indicators, Nearest Node Dynamic (NND) and Nearest Node Pretensioning (NNP), was evaluated across six seating configurations using both global and local coordinate systems. Validation, primarily based on the Matthews Correlation Coefficient (MCC), showed that NND has higher accuracy compared to NNP, mainly due to its dynamic node selection at each output step, unlike NNP's fixed node. The coordinate system choice had limited effect, with the global system sufficient even for reclined postures. Individual pelvic shape variations and the selected analysis period can influence indicator results and should be addressed in future studies. An extension of the NND indicator shows promise by enabling continuous rather than binary evaluation, potentially improving predictive accuracy.

Future research should expand testing with more diverse occupant models and leverage large-scale simulations, where automated and efficient submarining detection methods like these indicators are crucial. Overall, these findings support improved occupant safety and adaptable restraint system design for diverse passenger needs.

VI. REFERENCES

[1] Richardson, R., et al. (2020) Pelvis Kinematics and Injuries of Reclined Occupants in Frontal Impacts. *Proceedings of the IRCOBI Conference*, 2020, Online.

[2] Richardson, R., et al. (2019) Test Methodology for Evaluating the Reclined Seating Environment With Human Surrogates. *Proceedings of 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2019, Eindhoven, Netherlands.

[3] Richardson, R., et al. (2020) Kinematic and injury response of reclined PMHS in frontal impacts. *Stapp Car Crash Journal, 64,* pp.83–153.

[4] Jeong, G. and Kim, D. (2024) A Study of Submarining Occurrence Factors in Reclined Sitting Posture. *Proceedings of 68th Stapp Car Crash Conference*, 2024, Ohio, USA.

[5] Adomeit, D. and Heger, A. (1975) Motion sequence criteria and design proposals for restraint devices in order to avoid unfavorable biomechanic conditions and submarining. *SAE Transactions*, **84**: pp. 3150–3159.

[6] Ji, P., Huang, Y., Zhou, Q. (2017) Mechanisms of using knee bolster to control kinematical motion of occupant in reclined posture for lowering injury risk. International Journal of Crashworthiness, **22**(4), pp. 415–424.

[7] Kitagawa, Y., Hayashi, S., Yamada, K., Gotoh, M. (2017) Occupant Kinematics in Simulated Autonomous Driving Vehicle Collisions: Influence of Seating Position, Direction and Angle (No. 2017-22-0005). SAE Technical Paper.

[8] Rawska K., Gepner B., Moreau D., Kerrigan JR. (2020) Submarining sensitivity across varied seat configurations in autonomous driving system environment. *Traffic Injury Prevention*, **21**(sup1): pp. S1-S6.

[9] Gepner, B.D, et al. (2019) Comparison of Human Body Models in Frontal Crashes with Reclined. *Proceedings of the IRCOBI Conference*, 2019, Florence, Italy.

[10] Gepner, B.D, et al. (2020) Evaluation of GHBMC, THUMS and SAFER Human Body Models in Frontal Impacts in Reclined Postures. *Proceedings of the IRCOBI Conference*, 2022, Porto, Portugal.

[11] Ngo, A., Becker, J. (2021) Investigation of Occupant Kinematics and Injury Risk in a Reclined and Rearward-Facing Seat under Various Frontal Crash Velocities. *Journal of Safety Research,* **79**: pp. 26-37.

[12] Euro NCAP Vision 2030. Available at: https://cdn.euroncap.com/media/74468/euro-ncap-roadmap-vision-2030.pdf Accessed April 20, 2025.

[13] Boyle, K., Fanta, A., et al. (2020) Restraint systems considering occupant diversity and pre-crash posture. *Traffic Injury Prevention*, **21**(sup1): pp. S31–S36.
[14] Yang, Z., Desai, A., et al. (2023) Restraint System Optimizations Using Diverse Human Body Models in Frontal Crashes. *SAE International Journal of Transportation Safety*, **11**(2): pp. 187–196.
[15] Sun, W. et al. (2024) Efficient injury risk predictions for a diverse population using parametric human modeling and inducing points in Gaussian processes. *Traffic Injury Prevention*, **25**(sup1), pp. S43–S50
[16] Eggers, A., et al. (2023) The ENOP project (Enable New Occupant Seating Positions). *VDI-Berichte Fahrzeugsicherheit,* Band 2426, VDI-Verlag.
[17] Euro NCAP New Car Assessment Programme (2019) Adult Occupant Protection, v. 9.0.1. Available at: https://cdn.euroncap.com/media/43372/euro-ncap-assessment-protocol-aop-v901.pdf Accessed April 20, 2025.
[18] Pipkorn, B., et al. (2016) Submarining Evaluation of Belted Occupants in Frontal Crashes: A Comparison Between Human Body Models and ATDs. *Proceedings of the IRCOBI Conference*, 2016, Malaga, Spain.
[19] Richard, O., Lebarbé, M., et al. (2022) THOR-05F Response in Sled Tests Inducing Submarining and Comparison with PMHS Response Corridors. *SAE Technical Paper* 2021-22-0005.
[20] Luet, C., Trosseille, X., Drazétic, P., Potier, P., Vallancien, G. (2012) Kinematics and dynamics of the pelvis in the process of submarining using PMHS sled tests. *Stapp Car Crash Journal*, **56**: pp. 411–442.
[21] Shigeta, K., Kitagawa, Y., Yasiko, T. (2009) Development of next generation human body FE model capable of organ injury prediction. *Proceedings of 21th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, paper no. 09-0111, Stuttgart, Germany.
[22] Kitagawa, T. and Yasuki, Y. (2014) Development and Application of THUMS Version 4. Proceedings of CAHRS Human Modeling and Simulation in Automotive Engineering, p. 1–138, Munich, Germany.
[23] Wang, Z. and Zaseck L. (2022) THOR-AV 50th Percentile Male Biofidelity Evaluation in 25° and 45° Seatback Angle Test Conditions with a Semi-Rigid Seat. *Proceedings of the IRCOBI Conference*, 2022, Porto, Portugal.
[24] Uriot, J., et al. (2015) Reference PMHS Sled Tests to Assess Submarining. *Stapp Car Crash Journal,* vol. **59**, pp. 203-223.
[25] Trosseille, X., Petit, P., et al. (2018) Reference PMHS sled tests to assess submarining of the small female. *Stapp Car Crash Journal*, vol. **62**, pp. 93-118.
[26] NTHSA (2022) Integrated Seat Belt System Model Development. Available at: https://www.nhtsa.gov/crash-simulation-vehicle-models Accessed April 20, 2025.
[27] Peres, J., Praxl, N. (2023) A Simulation Study on the Knee-Thigh-Hip Loading of the THOR Compared to Human Body Models. *Proceedings of the 27th International Technical Conference on the Enhanced Safety of Vehicles (ESV*), 2018, Yokohama, Japan.
[28] Huf, A., et al. (2018) Development of Occupant Restraint Systems for Future Seating Positions in Fully or Semi Autonomous Vehicles. *Proceedings of Airbag Conference*, 2018, Chennai, India.
[29] Atanassov, E. and Ivanovska, S. (2022) On the Use of Sobol Sequence for High Dimensional Simulation. In: *Computational Science* – ICCS 2022, 2022 June: pp. 646–652.
[30] Thums User Community (TUC). Available at: https://tuc-project.org/ Accessed April 20, 2025.
[31] Wilcoxon, F., (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, vol. **1**, no. 6, 1945, pp. 80–83.
[32] Dexter, F., (2013) Wilcoxon-Mann-Whitney test used for data that are not normally distributed. *Anesth Analg*. **117**(3):537-538.
[33] Foody, GM., (2023) Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PloS one* vol. **18**,10
[34] Chicco D, Jurman G., (2023) The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min*., **16**(1):4.
[35] Chicco, D., Jurman, G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6.
[36] Kulkarni, A., Chong, D., Batarseh, F. A. (2020) Foundations of data imbalance and solutions for a data democracy. *In Elsevier eBooks* (pp. 83–106).

[37] Holt, C. A., Sullivan, S. P. (2023) Permutation tests for experimental data. *Experimental Economics*, **26**(4), 775–812.
[38] Izumiyama, I., et al. (2022) Analysis of Individual Variabilities for Lumbar and Pelvic Alignment in Highly Reclined Seating Postures and Occupant Kinematics in a Collision. *Proceedings of the IRCOBI Conference*, 2022, Porto, Portugal.
[39] Rees, M., et al. (2023) Effects of Geometric Variations of Anterior Iliac Wings on Submarining Risk in Reclined Seating. *Proceedings of the IRCOBI Conference*, 2023, Cambridge, UK.
[40] Richardson, R. E., Gepner, B. D., Kerrigan, J. R., Forman, J. L. (2024) Evaluation of lap belt-pelvis load transfer in frontal impact simulations using finite element occupant models. *Proceedings of the Traffic Injury Prevention*, **25**(8), 1137–1145.
[41] Naseri H, Iraeus J, Johansson H., (2022) A numerical study on the safety belt-to-pelvis interaction. *International journal for numerical methods in biomedical engineering,* **38**(4):e3572.
[42] Saito, H., Pipkorn, B., Lubbe, N., (2022) Understanding the Influence of Seat Belt Geometries on Belt-to-Pelvis Angle Can Help Prevent Submarining. *SAE International Journal of Transportation Safety*, **10**(2):463-481.

VII. APPENDIX

Appendix A
Pair-Plot of Sobol Sampling

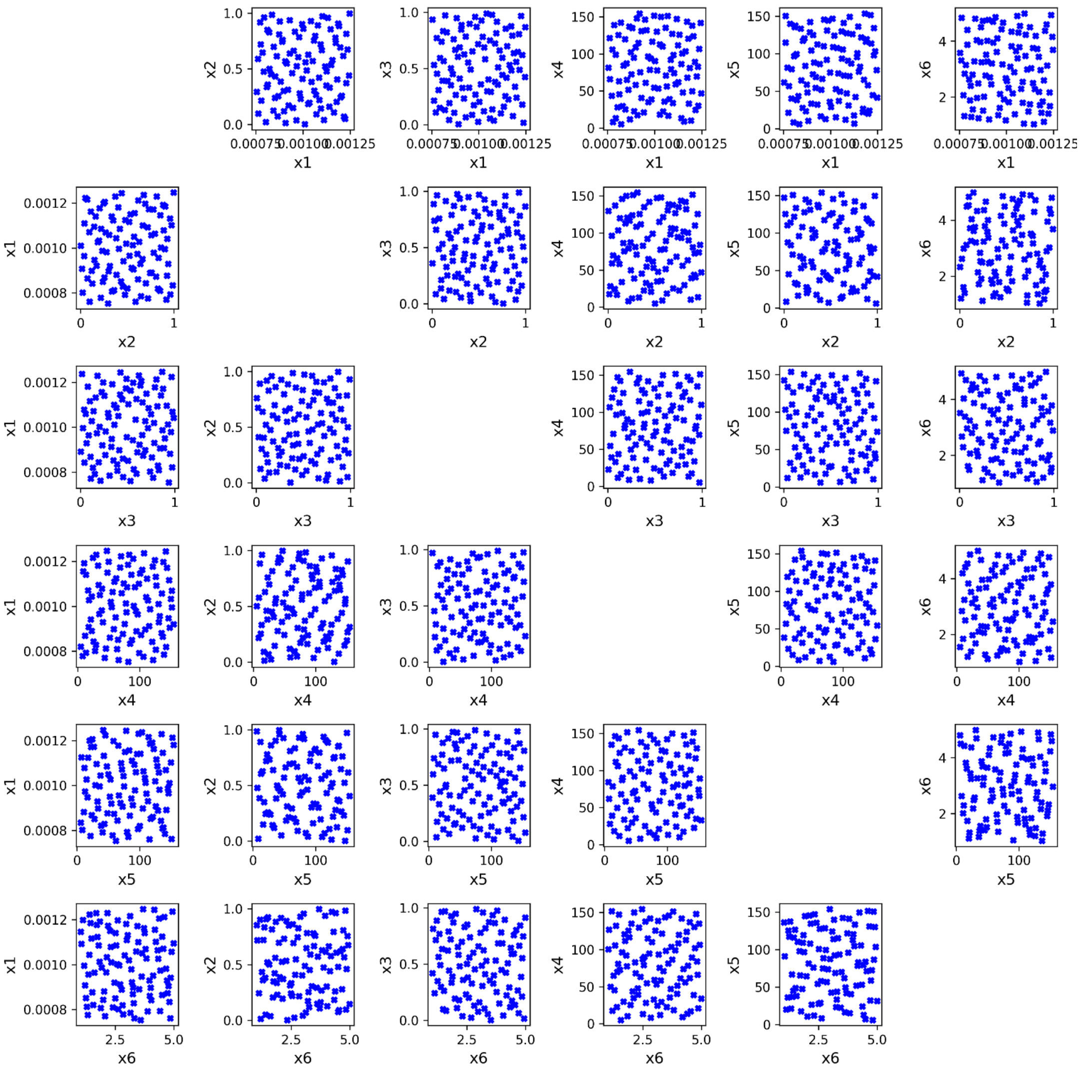

Fig. A1. Sobol Sampling of all varied input parameter.

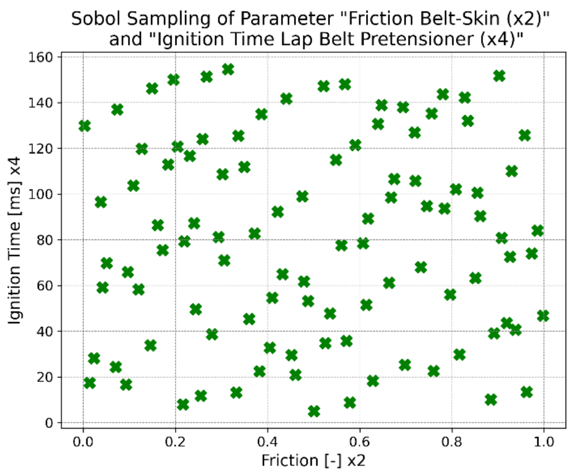

Fig. A2. Sobol Sampling for one parameter combination from Fig. A1, with friction x2 on the x-axes and pretensioner ignition time x4 on the y-axes.

Appendix B

Example of one simulation in Tandem [C] environment at four time steps for a submarining case

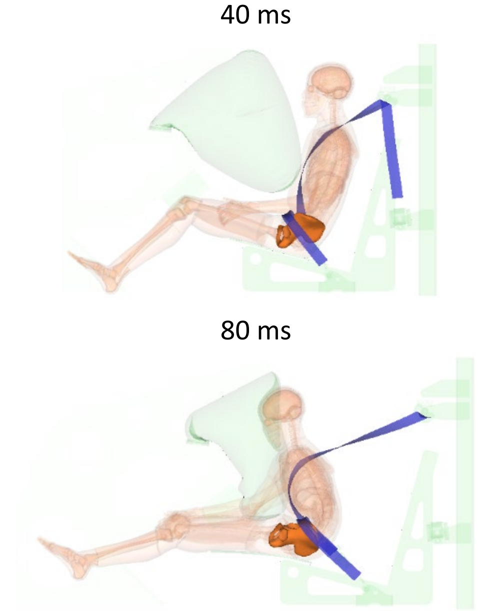

Fig. B1. THUMS v4.1 50th in Tandem sled [C], simulation 01/100 with submarining occurrence on left and right ASIS side.

Appendix C

Example of one simulation in [B2] environment at four time steps for a submarining case

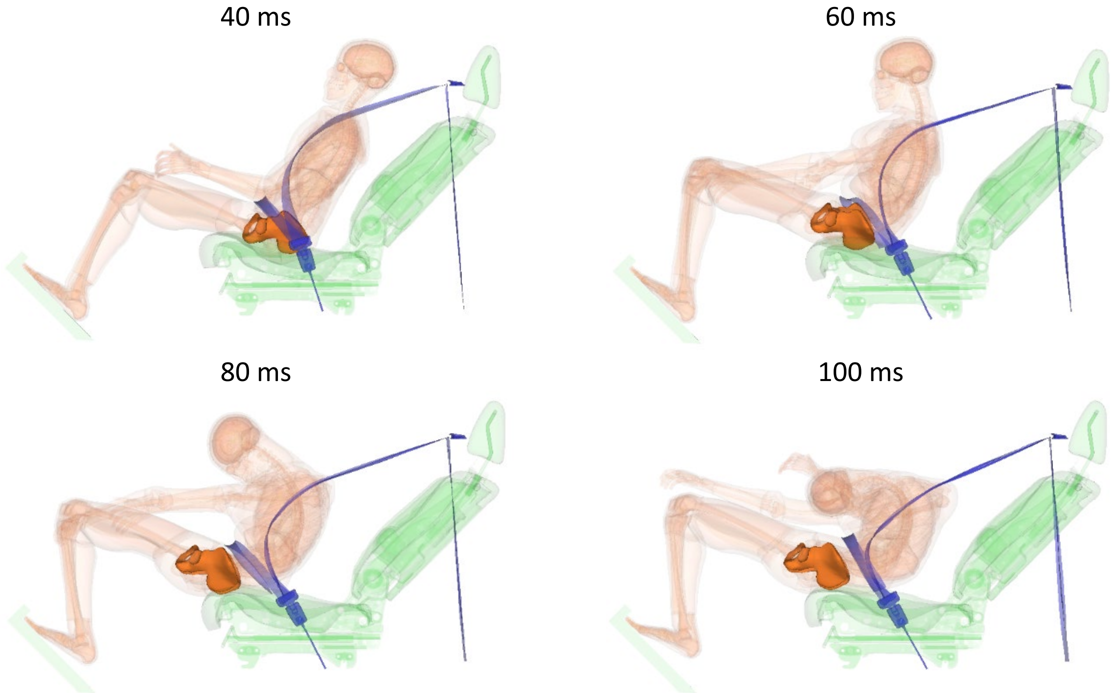

Fig. C1. THUMS v4.1 50th in LAB/CEESAR-Integrated NHTSA Seat [B2], simulation 21/100 with submarining occurrence on left and right ASIS side.

Appendix D

Results of the Wilcoxon rank-sum test in baseline and validation study: log-transformed p-values for all four indicators, shown separately for left and right ASIS sides across all six environments.
***Left*: baseline study (4 indicators × 2 sides × 6 environments = 48 curves)**
***Right*: validation study (4 x 2 x 2 = 16 curves)**

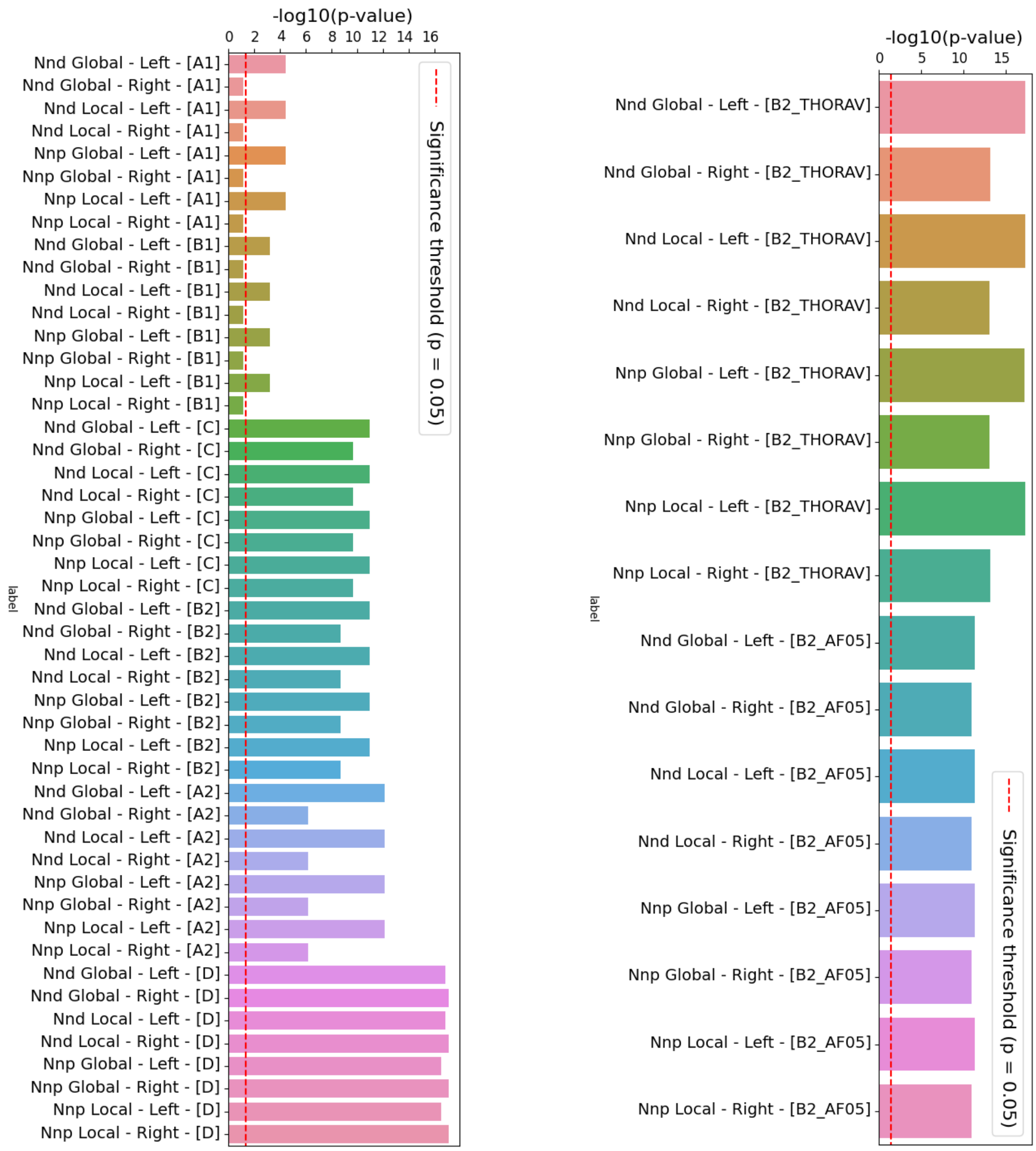

Fig. D1. *Left*: p-values (log) for baseline study. *Right*: p-values (log) for validation study. Dashed line indicates the significance threshold (p = 0.05).

Appendix E - Indicator results for all environments in the baseline study

	Load case	Submarining Cases	Non-Submarining Cases
[A1]	LAB/CEESAR semi-rigid seat (Upr)	6	94

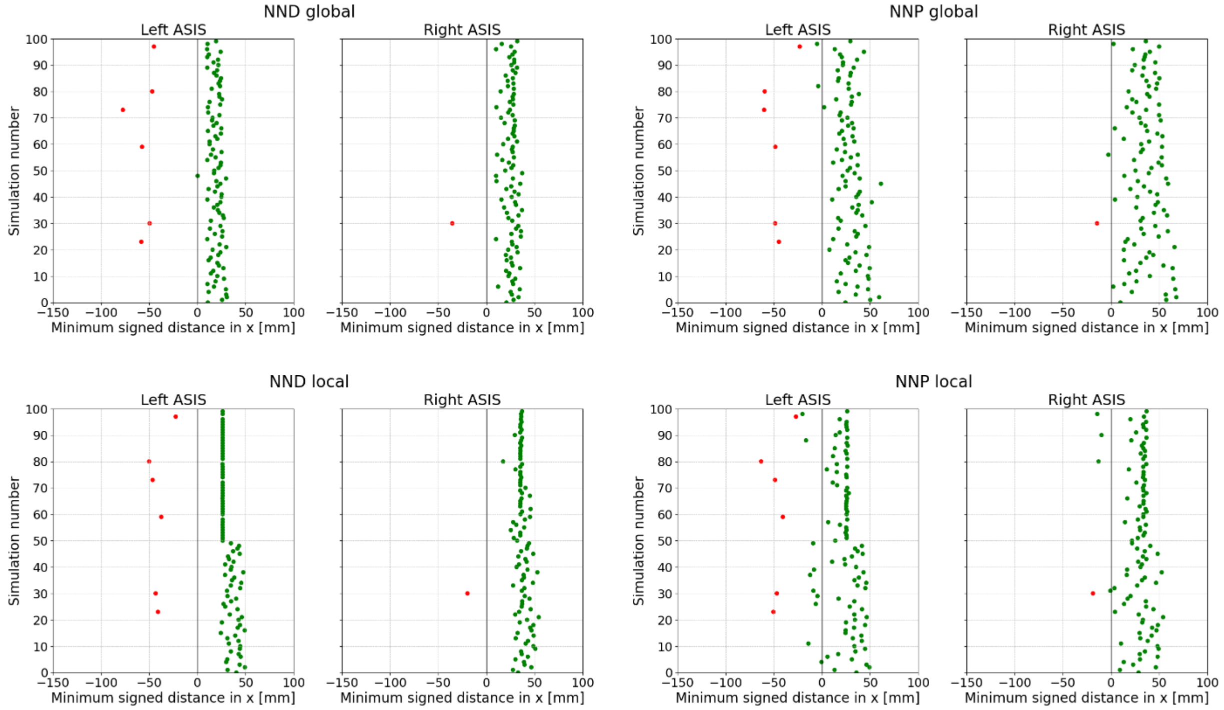

Fig. E1. Comparison of indicator predictions to true label information for configuration [A1].
NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

	Load case	Submarining Cases	Non-Submarining Cases
[B1]	LAB/CEESAR-Integrated NHTSA Seat	4	96

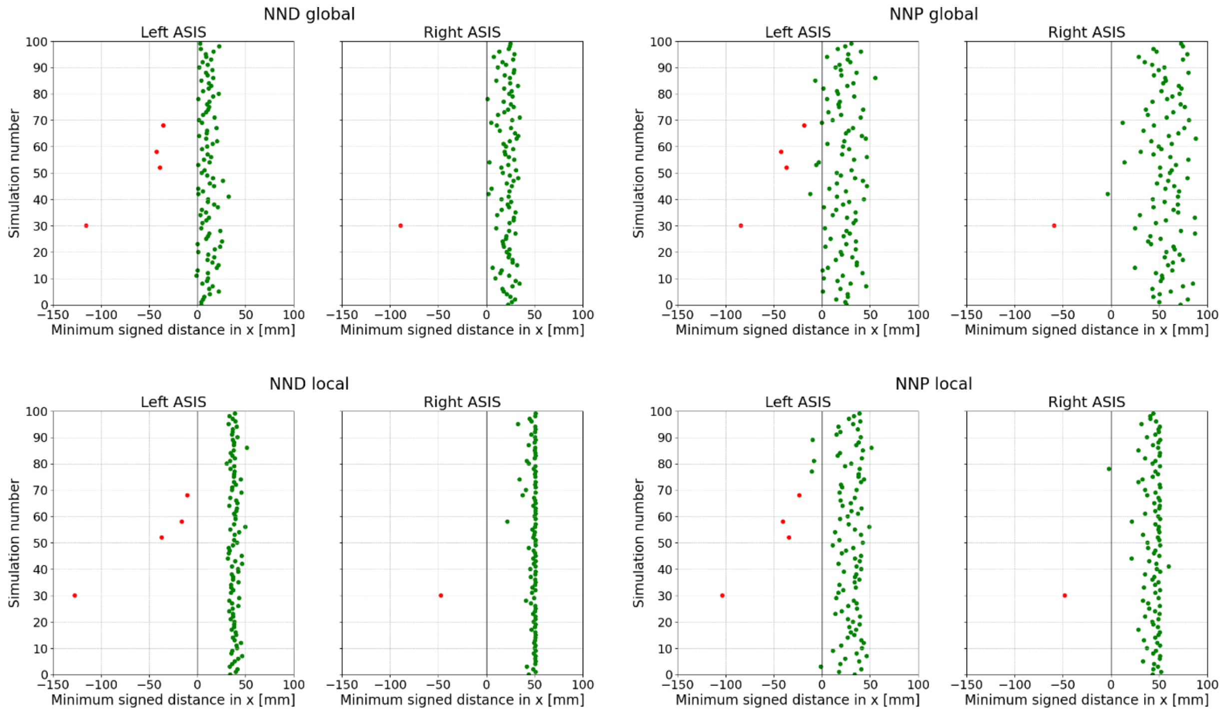

Fig. E2. Comparison of indicator predictions to true label information for configuration [B1].
NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

	Load case	Submarining Cases	Non-Submarining Cases
[C]	Tandem Sled	19	81

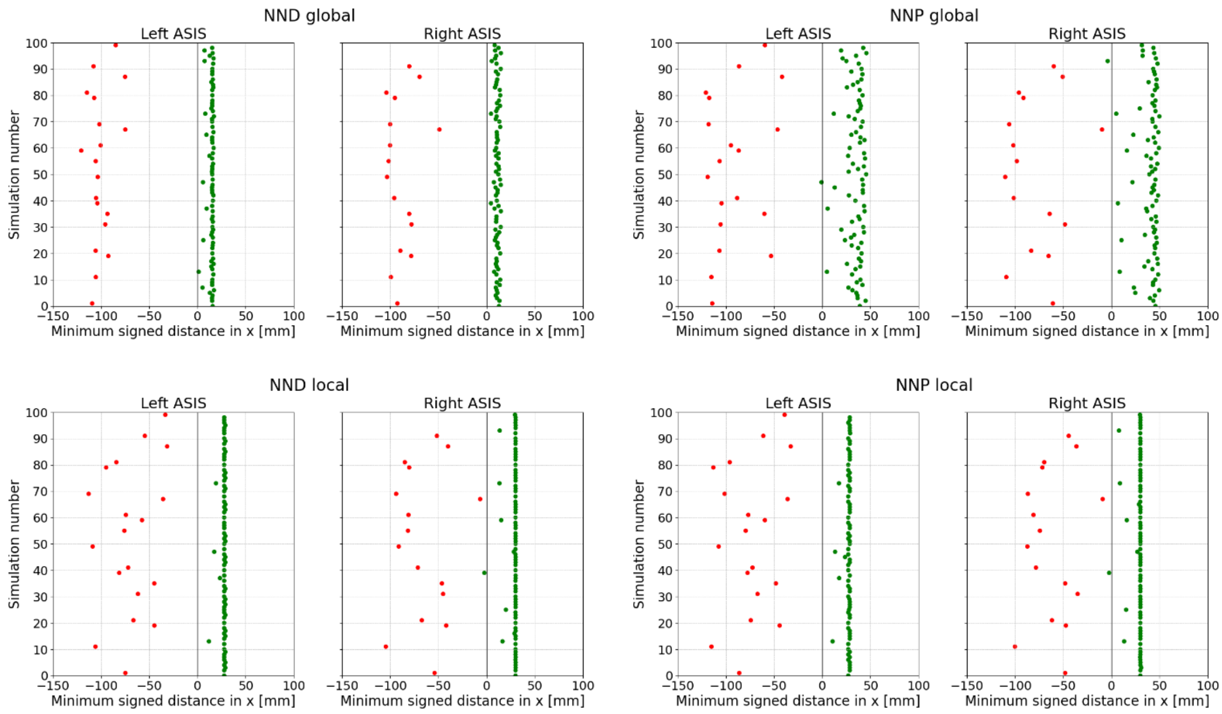

Fig. E3. Comparison of indicator predictions to true label information for configuration [C].
NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

	Load case	Submarining Cases	Non-Submarining Cases
[A2]	LAB/CEESAR semi-rigid seat (Recl)	22	78

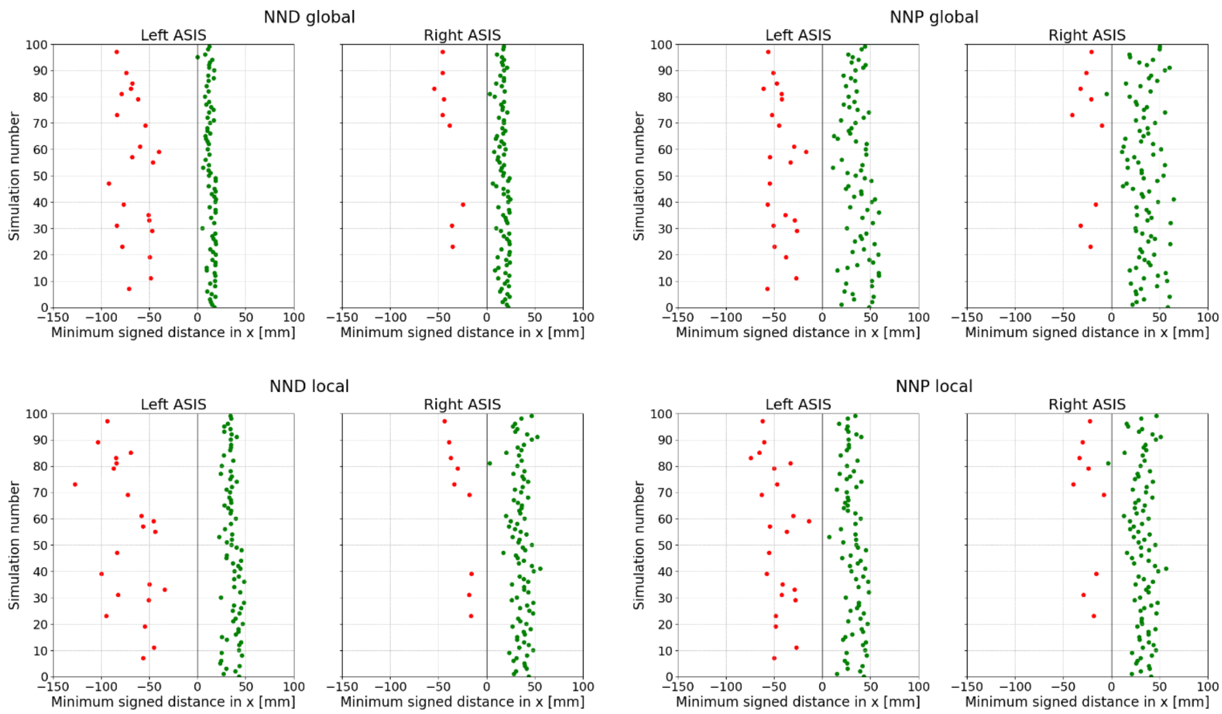

Fig. E4. Comparison of indicator predictions to true label information for configuration [A2].
NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

	Load case	Submarining Cases	Non-Submarining Cases
[B2]	LAB/CEESAR-Integrated NHTSA Seat	19	81

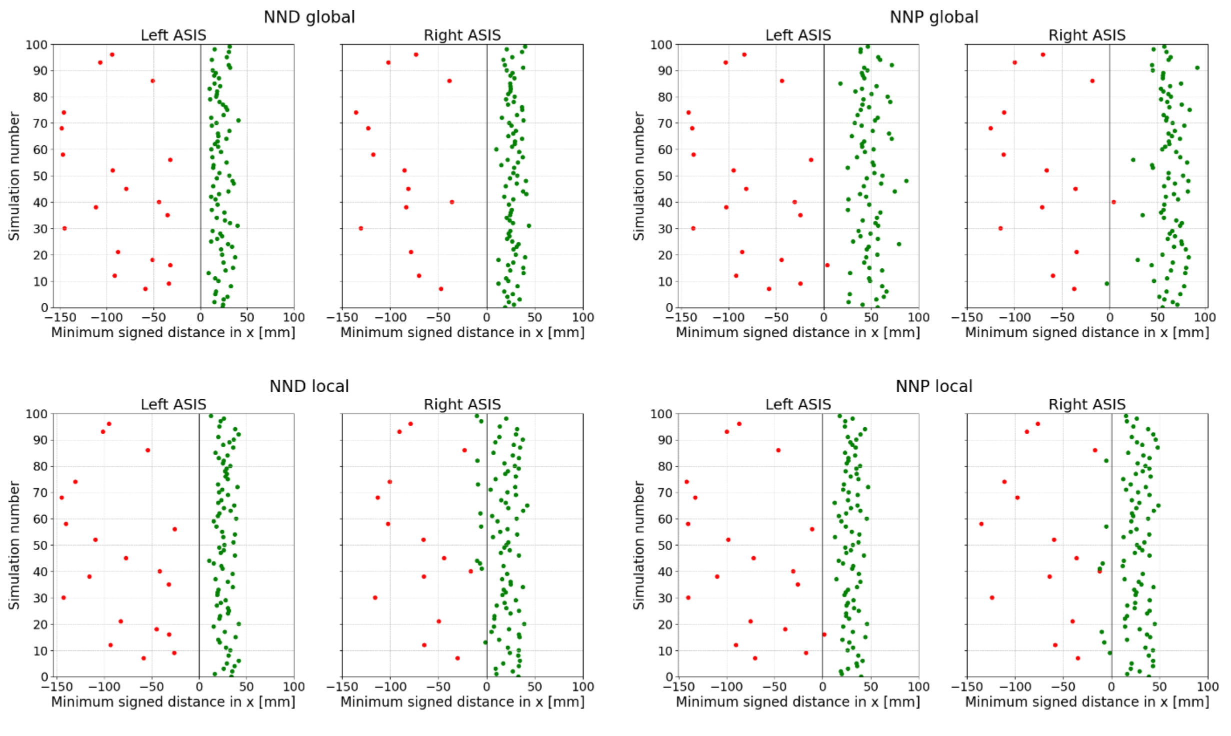

Fig. E5. Comparison of indicator predictions to true label information for configuration [B2].
NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

	Load case	Submarining Cases	Non-Submarining Cases
[D]	Comfort Seat	58	42

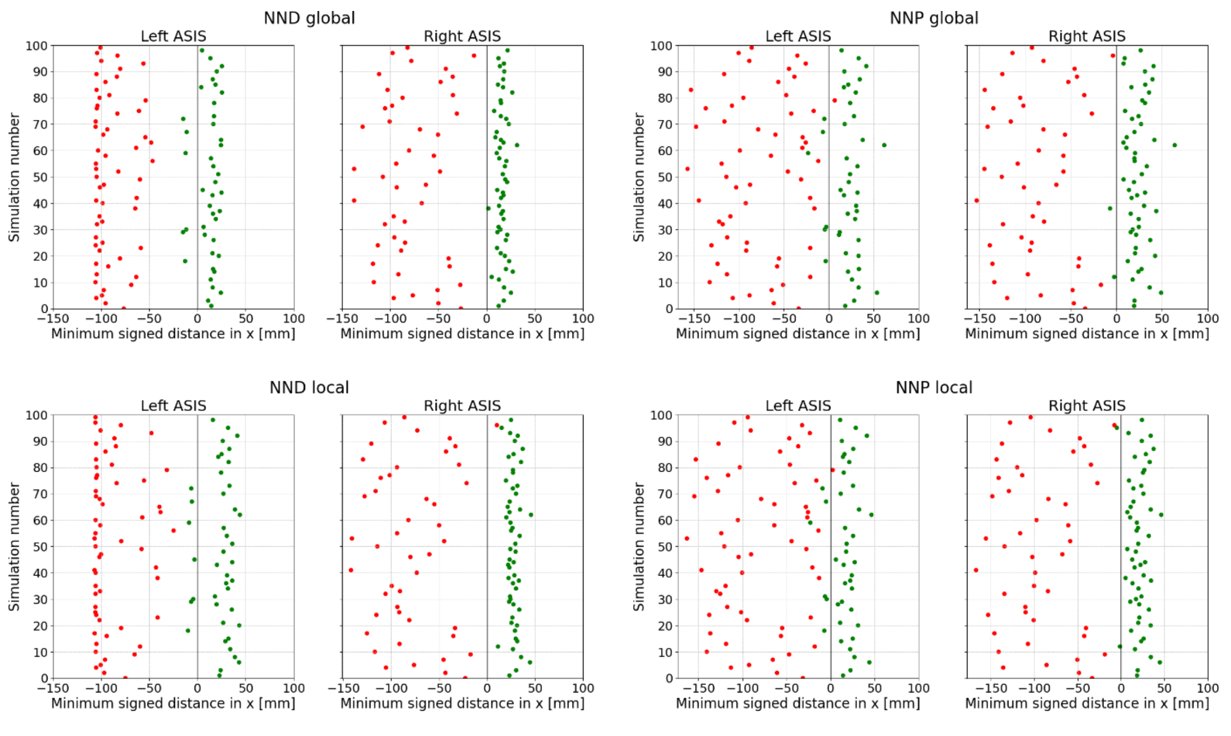

Fig. E6. Comparison of indicator predictions to true label information for configuration [D].
NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

In the upright posture scenario, six submarining and 94 non-submarining cases were identified. Figures Fig. E1-Fig. E6 summarise the indicator outputs. The left column shows NND results (top: global, bottom: local), and the right column shows NNP outputs. All submarining cases show negative values. Non-submarining cases cluster near zero, with some exceeding it in both NNP variants. This is not seen in NND. Three of four plots show clear separation between submarining and non-submarining, especially both NND and global NNP. Local NNP shows less separation. Local evaluations in most cases shift values toward the positive, reduce scatter, and compress distribution. For instance, run 98 (left ASIS) in NND shifts from -55 mm (global) to -25 mm (local).

In the second environment, a similar pattern appears: submarining cases are negative, non-submarining near or above zero. For the left ASIS, NND separates better than NNP due to higher negative values in submarining cases. NNP shows higher dispersion, especially above zero. Local evaluations show reduced scatter, especially on the right. Local NND values shift toward zero.

The Tandem sled, also with upright posture, shows submarining in 20% of cases, caused by variable conditions. Both indicators accurately separate cases. In NNP, non-submarining values reach +50 mm, while in NND they stay below +25 mm. In local evaluations (more pronounced in NND as NNP), separation weakens due to rightward shift in values, stronger for negative than positive values. For example, run 67 (right ASIS) shifts from -50 mm (global) to -5 mm (local); run 0 shifts from +12 mm to +31 mm. All 19 (left) and 16 (right) submarining cases are correctly classified by all indicators.

In the first reclined posture, separation is visible for all indicators, though varying in extent. NND values are more dispersed than NNP, especially on the left ASIS.

In environment [D], submarining occurs most often, and the separation between classes is weaker. The global NND indicator shows the clearest distinction, although one submarining case is not detected, and some non-submarining cases appear in the negative range. This underscores the importance of the threshold, which is currently set at 0. The area around this threshold includes several borderline cases. For global NND, clear submarining cases only appear around -50, while the range between -50 and 0 includes uncertain outcomes. Local indicators show less clear separation, as their values lie closer together. In both local variants, the ability to distinguish between classes is reduced.

Appendix F – Indicator results for both occupant models in validation study for environment [B2]

	Load case	Submarining Cases	Non-Submarining Cases
[B2]	LAB/CEESAR-Integrated NHTSA Seat – THUMS v4.1 05th	20	80

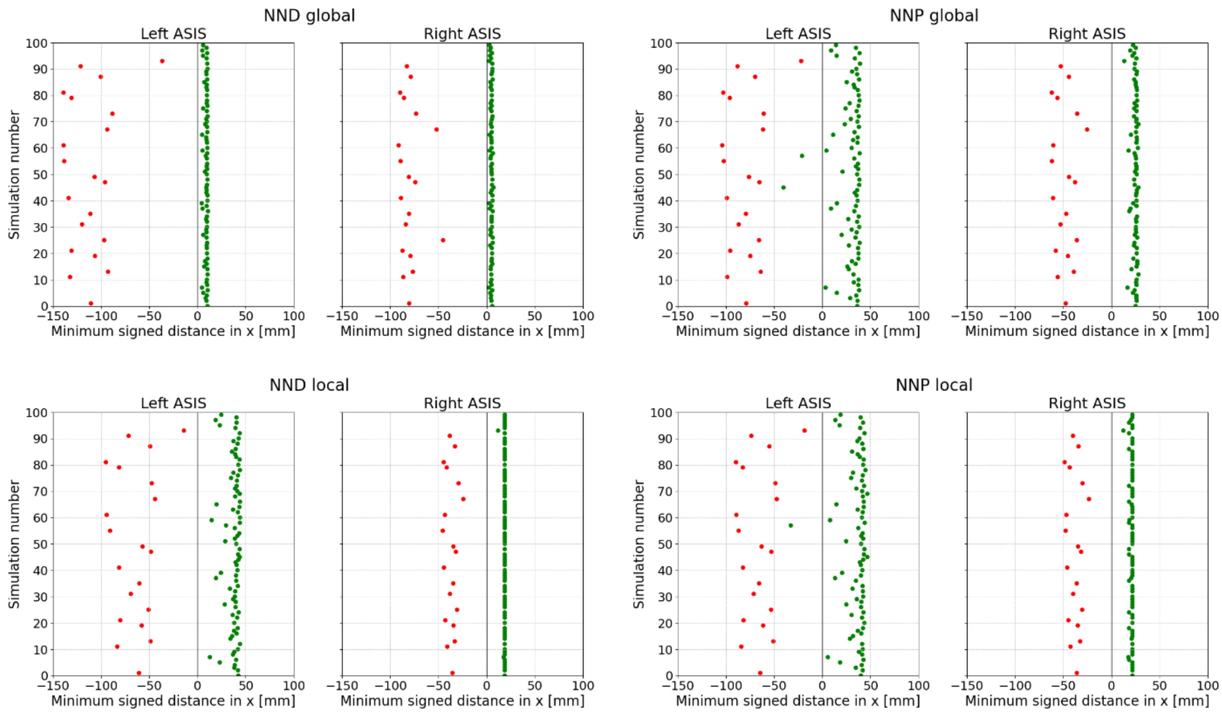

Fig. F1. Comparison of indicator predictions to true label information for configuration [B2] with THUMS AF05. NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

	Load case	Submarining Cases	Non-Submarining Cases
[B2]	LAB/CEESAR-Integrated NHTSA Seat – THOR AV 50M	50	50

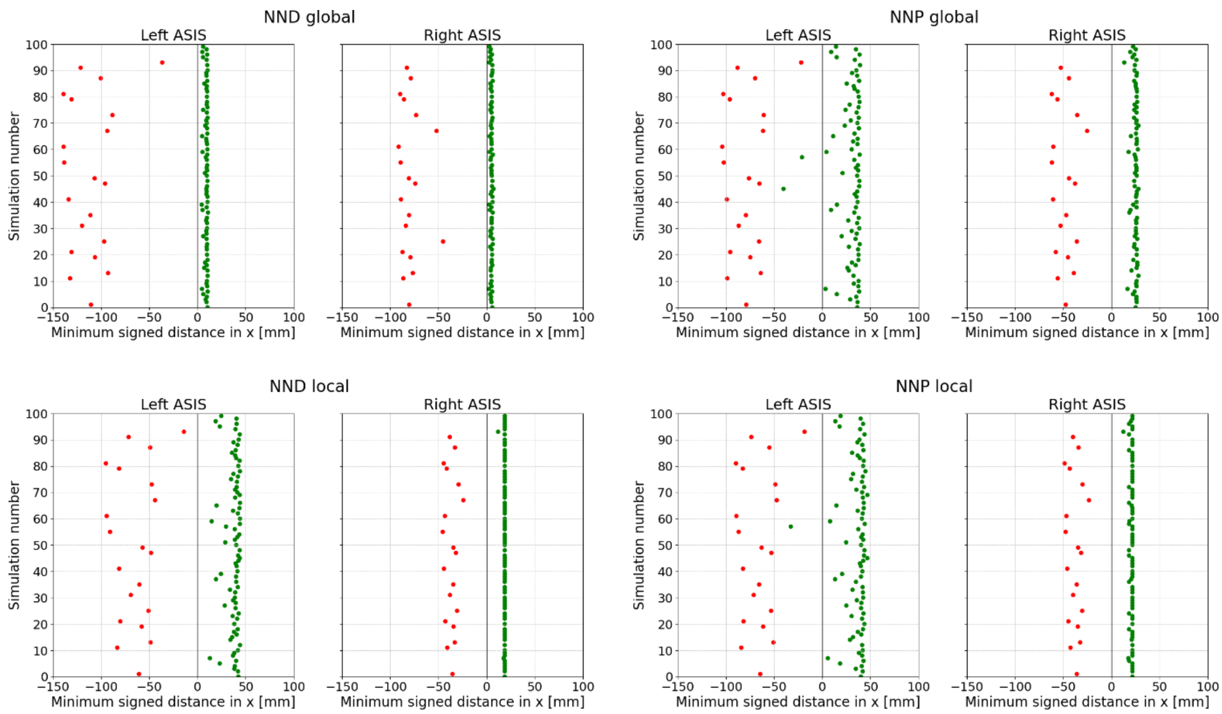

Fig. F2. Comparison of indicator predictions to true label information for configuration [B2] with THOR AV 50M. NND globally (*top left*), locally (*bottom left*), NNP globally (*top right*), NNP locally (*bottom right*).

Appendix G - Confusion Matrix and related Matthews-Correlation-Coefficients for all environments in the baseline study

The Matthews Correlation Coefficient (MCC) was chosen as the primary metric because it captures the overall quality of binary classifications by considering all elements of the confusion matrix (TP, TN, FP, FN). Unlike metrics such as Recall or Precision, MCC does not favor one type of misclassification over another, making it especially suitable for imbalanced datasets like those in this study.

Its balanced nature enables a robust performance comparison across diverse seating environments. Averaging MCC across the six environments provides a concise and meaningful summary of indicator performance. This approach is appropriate here as no individual MCC value was close to zero, i.e. no indicator completely failed in any environment, which could otherwise skew the average.

While MCC offers a single, balanced value, other metrics can highlight specific error tendencies (like Recall for FN sensitivity, Precision for FP). For completeness, these additional metrics are reported and discussed in Appendix I.

TABLE G.I
CONFUSION MATRIX FOR ENVIRONMENT [A1] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT

Environment [A1]		TP	TN	FP	FN	MCC	Ø MCC
NND global	Left	6	94	0	0	1	1
	Right	1	99	0	0	1	
NND local	Left	6	94	0	0	1	1
	Right	1	99	0	0	1	
NNP global	Left	6	92	2	0	0.86	0.78
	Right	1	98	1	0	0.70	
NNP local	Left	6	84	10	0	0.58	0.51
	Right	1	95	4	0	0.44	

TABLE G.II
CONFUSION MATRIX FOR ENVIRONMENT [B1] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT

Environment [B1]		TP	TN	FP	FN	MCC	Ø MCC
NND global	Left	4	96	0	0	1	1
	Right	1	99	0	0	1	
NND local	Left	4	96	0	0	1	1
	Right	1	99	0	0	1	
NNP global	Left	4	91	5	0	0.65	0.67
	Right	1	98	1	0	0.70	
NNP local	Left	4	92	4	0	0.69	0.69
	Right	1	98	1	0	0.70	

TABLE G.III
CONFUSION MATRIX FOR ENVIRONMENT [C] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT

Environment [C]		TP	TN	FP	FN	MCC	Ø MCC
NND global	Left	19	81	0	0	1	1
	Right	16	84	0	0	1	
NND local	Left	19	81	0	0	1	0.98
	Right	16	83	1	0	0.96	
NNP global	Left	19	80	1	0	0.97	0.97
	Right	16	83	1	0	0.96	
NNP local	Left	19	81	0	0	1	0.98
	Right	16	83	1	0	0.96	

TABLE G.IV

CONFUSION MATRIX FOR ENVIRONMENT [A2] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT

Environment [A2]		TP	TN	FP	FN	MCC	Ø MCC
NND global	Left	22	78	0	0	1	1
	Right	9	91	0	0	1	
NND local	Left	22	78	0	0	1	1
	Right	9	91	0	0	1	
NNP global	Left	22	78	0	0	1	0.97
	Right	9	90	1	0	0.94	
NNP local	Left	22	78	0	0	1	0.97
	Right	9	90	1	0	0.94	

TABLE G.V

CONFUSION MATRIX FOR ENVIRONMENT [D] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT

Environment [D]		TP	TN	FP	FN	MCC	Ø MCC
NND global	Left	58	36	6	0	0.88	0.94
	Right	46	55	0	0	1	
NND local	Left	58	35	7	0	0.86	0.92
	Right	45	54	0	1	0.98	
NNP global	Left	57	36	6	1	0.86	0.91
	Right	46	52	2	0	0.96	
NNP local	Left	57	36	6	1	0.86	0.91
	Right	46	52	2	0	0.96	

Appendix H
Confusion Matrix and related Matthews-Correlation-Coefficients for both occupant models in the validation study in environment [B2]

TABLE H.I

CONFUSION MATRIX FOR ENVIRONMENT [B2] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT – THUMS V4.1 05TH

Indicators		TP	TN	FP	FN	MCC	Ø MCC
NND global	Left	20	80	0	0	1	1
	Right	19	81	0	0	1	
NND local	Left	20	80	0	0	1	1
	Right	19	81	0	0	1	
NNP global	Left	20	78	2	0	0.941	0.97
	Right	19	81	0	0	1	
NNP local	Left	20	79	1	0	0.91	0.95
	Right	19	81	0	0	1	

TABLE H.II

CONFUSION MATRIX FOR ENVIRONMENT [B2] AND ALL FOUR INDICATORS, INCLUDING THE MC-COEFFICIENT – THOR AV 50M

Indicators		TP	TN	FP	FN	MCC	Ø MCC
NND global	Left	50	46	4	0	0.96	0.96
	Right	25	75	0	0	1	
NND local	Left	48	50	0	2	0.96	0.94
	Right	23	74	1	2	0.92	
NNP global	Left	45	50	0	5	0.90	0.92
	Right	23	75	0	2	0.97	
NNP local	Left	44	50	0	6	0.88	0.92
	Right	24	75	0	1	0.97	

Appendix I
Additional Classification Metrics: Recall, Precision, F1-Score for all setups in baseline study

In addition to MCC, three commonly used classification metrics, Recall, Precision, and F1-Score, were considered to provide a more differentiated view of indicator performance. These metrics are particularly relevant when dealing with imbalanced datasets, as is the case in this study, where submarining events are much less frequent than non-submarining ones.

Recall (also known as sensitivity or true positive rate) measures the proportion of actual submarining cases that were correctly identified. It is especially important in safety-critical contexts where missing a true submarining event (false negative) should be avoided. For example in occupant safety, a high recall ensures that critical cases are not overlooked, even if this increases the number of false alarms.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

Precision measures the proportion of predicted submarining cases that were actually correct. It becomes relevant when false positives are costly or misleading, such as when an engineer is prompted to investigate a simulation flagged as submarining, but it turns out to be safe.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

F1-Score is the harmonic mean of Precision and Recall and provides a balanced view when both false positives and false negatives are relevant. It is especially useful when the dataset is imbalanced and no clear priority exists between minimising FN or FP.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{4}$$

While all three metrics are suitable for imbalanced data, they differ in what they emphasise. Recall favors sensitivity, Precision favors specificity, and F1-Score offers a compromise. MCC, by contrast, considers all four confusion matrix components (TP, TN, FP, FN) and yields a single, balanced score. MCC can be less intuitive to interpret, which makes these additional metrics a useful complement for practical evaluation. Therefore the results for the three additional metrics are summarised in the following Figures Fig. I1, Fig. I2 and Fig. I3.

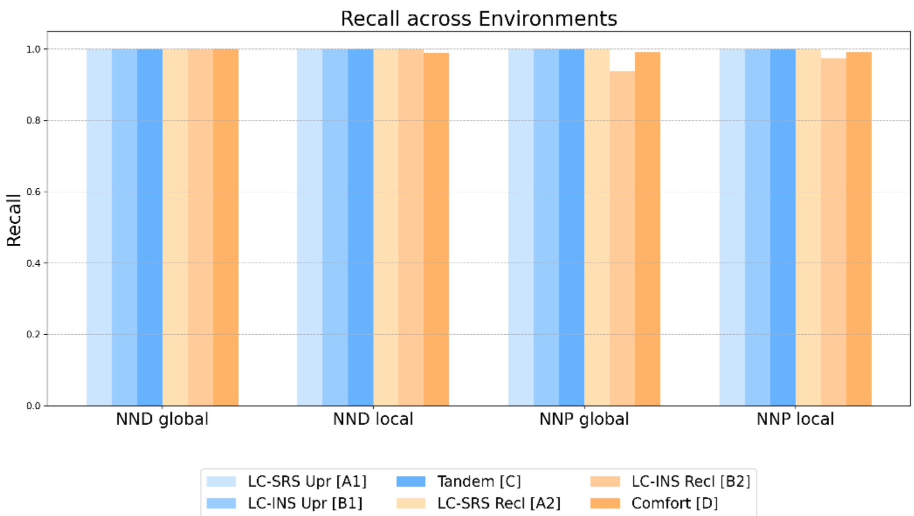

Fig. I1. Recall-Metric evaluated in baseline study across all six environments.

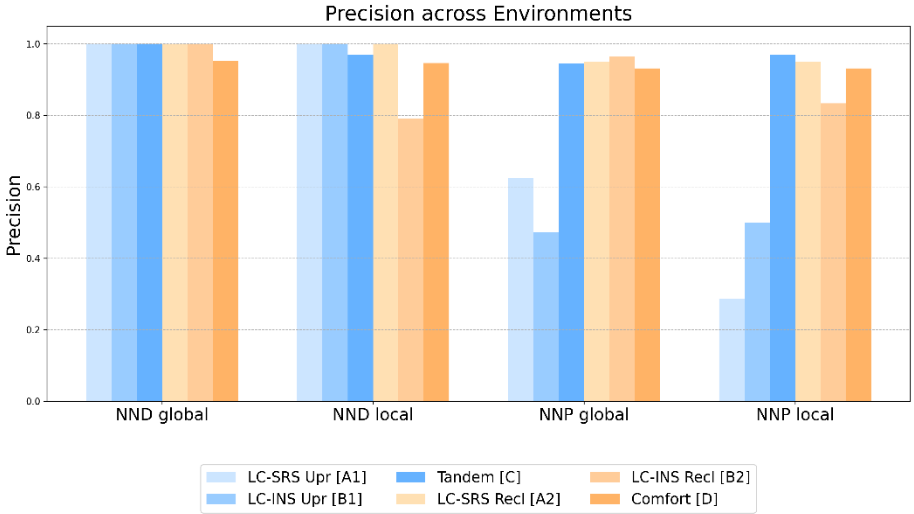

Fig. I2. Precision-Metric evaluated in baseline study across all six environments.

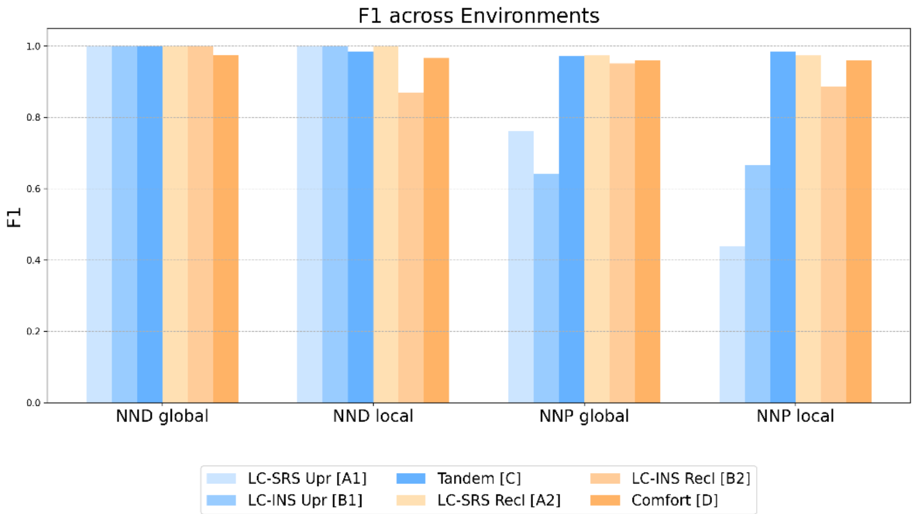

Fig. I3. F1-Score-Metric evaluated in baseline study across all six environments.

The recall plot demonstrates uniformly high values across all indicators, with nearly all reaching 1.0 in each environment. The lowest observed value, 0.94 for NNP global in [B2], still indicates a very high sensitivity. This suggests that all indicators reliably identify actual positive instances with minimal omission errors. In contrast, the precision plot reveals greater variation between indicators. NND global and NND local consistently achieve high or maximum precision values across all environments. NNP local exhibits substantial declines, particularly in [A1] (0.28) and [B1] (0.50), indicating a comparatively higher false positive rate. Even in environments with higher scores, such as [C], NNP local does not exceed 0.97, while NND global maintains 1.0. These differences are reflected in the F1-score, which balances recall and precision. NND global consistently achieves an F1-score of 1.0 in four out of six environments and shows the most stable performance across all scenarios. In contrast, NNP local and NNP global show marked reductions in environments with lower precision, with values as low as 0.44 and 0.76, respectively, in [A1]. In summary, while all indicators demonstrate high recall, NND local and NND global provide more balanced and consistently high performance across all three metrics. Their stability across diverse conditions underscores their robustness and suggests they are less affected by overestimation tendencies compared to the NNP indicators. These findings are consistent with the results from the MCC evaluation and support the observation that NND-based approaches tend to perform more consistently across metrics.

Treshold optimization:

So far, the threshold was set heuristically to 0 mm, but this does not guarantee optimal separation between submarining and non-submarining cases. To address this, threshold optimization can be performed by systematically varying the decision boundary and calculating a performance metric (e.g. F1-score or MCC) at each step. The threshold that maximizes the selected metric is then chosen as the optimal decision point. An initial example of such an approach is shown in Fig. X (Appendix X), where the F1-score improves from 0.88 to 1.0 by adjusting the threshold from 0 mm to –30 mm. This result is preliminary and intended as a first demonstration of the method and may be pursued further in future studies.

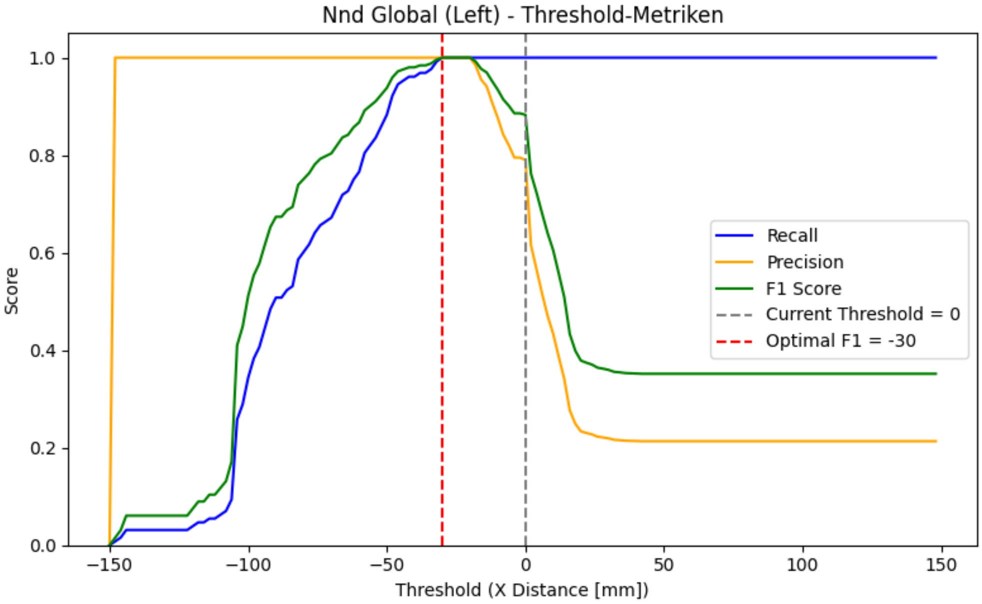

Fig. I4. Example of threshold optimization based on F1 score for a single indicator output.

Appendix J
Classification Metrics MCC, Recall, Precision and F1-Score for all setups in validation study

For the THOR AV 50M, the NND global indicator achieves the highest MCC (0.96), reflecting strong predictive capability despite some false positives. NND local follows closely (0.94), while both NNP indicators show slightly lower MCC values around 0.93, suggesting relatively reduced discriminative power. Results for the THUMS v4.1 05th show that all NND-based indicators reach the highest MCC of 1.0, indicating an ideal classification. NNP indicators also perform strongly (0.98 and 0.97), but minor false positives prevent perfect scores. These results highlight the more stable and consistent performance of NND-based indicators when evaluated by MCC.

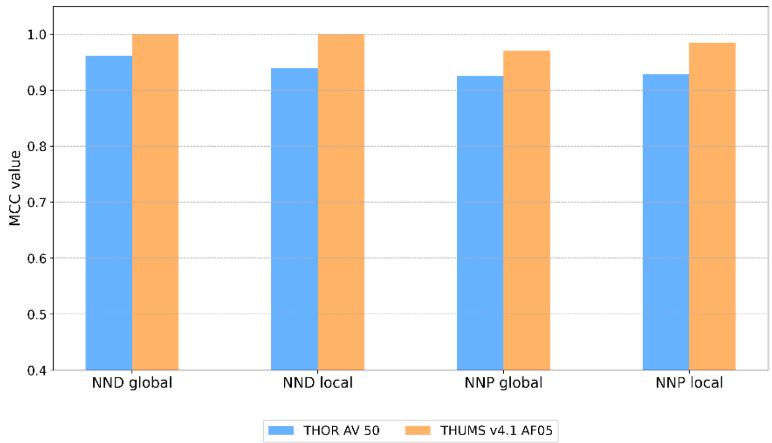

Fig. J2. MCC-Metric evaluated in validation study for both occupant models.

Analysing Recall, Precision, and F1-Score (Fig. J3, Fig. J4, Fig. J5) provides additional detail beyond the MCC by distinguishing between types of classification errors. Recall values are consistently high across all indicators, with NND global achieving perfect recall (1.0) in both surrogate models, indicating the highest sensitivity to true positives. The NNP indicators show slightly lower recall for the ATD runs (around 0.91-0.92), reflecting a small number of missed positive cases. Precision values are also very high for all indicators and both environments, mostly above 0.95. The only noticeable reduction occurs for NND global in the ATD scenario (0.96), caused by a few false positives. This confirms that false positive rates are low but differ slightly between indicators.

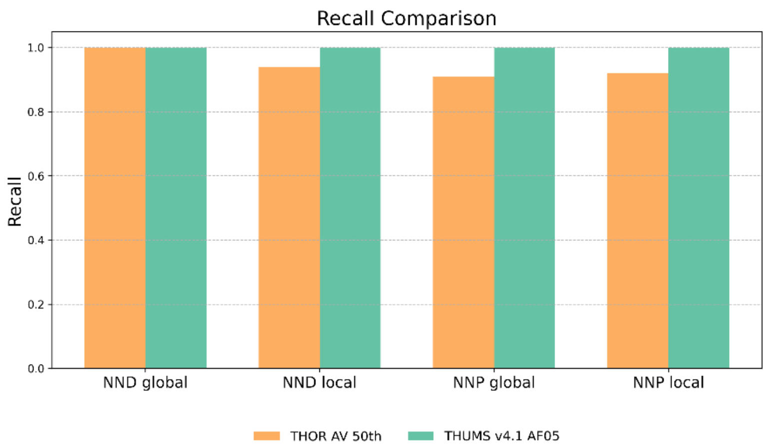

Fig. J3. Recall-Metric evaluated in validation study for both occupant models.

The F1-Score, which balances recall and precision, shows a similar trend: NND global attains the highest values (0.98 and 1.0), indicating a well-balanced performance. NNP indicators have slightly lower F1 scores, consistent with their somewhat reduced recall. An important observation is the trade-off between recall and precision, especially in the ATD runs, where both false negatives and false positives affect the indicators differently. Interpreting these metrics requires considering the specific context of application, in terms of whether missing positive cases or generating false alarms is more critical. Overall, the NND-based indicators demonstrate a moderate but consistent advantage in balancing sensitivity and specificity, which aligns with the MCC results.

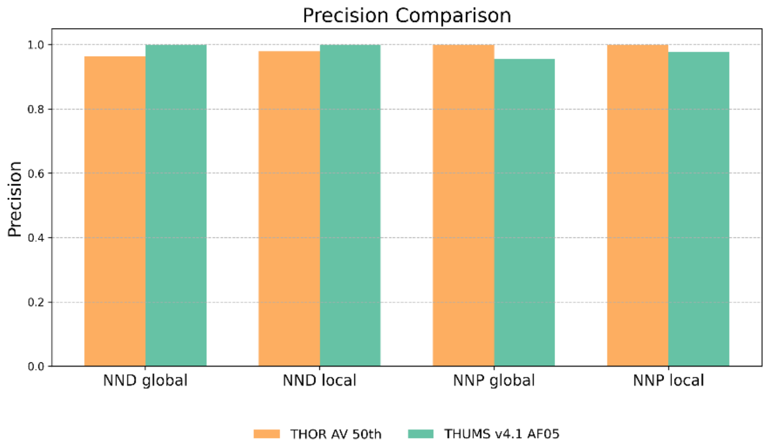

Fig. J4. Precision-Metric evaluated in validation study for both occupant models.

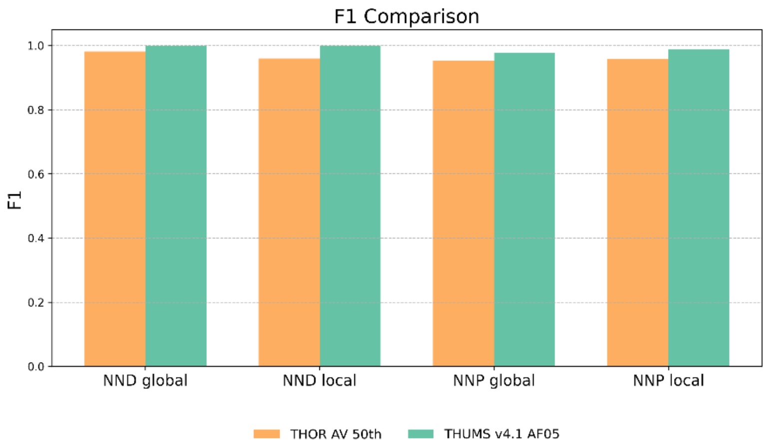

Fig. J5. F1-Score-Metric evaluated in validation study for both occupant models.

Appendix K
Updated Indicator definition

The following section describes an extension of the NND/NNP indicator. It is important to mention here that these are preliminary results that require further validation and investigation. More detailed findings will be presented in a subsequent publication. In addition to the distance measurement based on NND/NNP calculations already shown in this study, the maximum z-distance between ASIS and nearest belt node is scaled relative to the initial webbing width. This yields a percentage-based coverage value that indicates how close or far a configuration is from a submarining scenario. Fig. K1 shows a non-submarining case, in which the measured z-distance between the ASIS and the furthest belt node always remains positive. The scaled coverage value does not drop to 0%. The minimum coverage reaches approximately 45% on the right side, meaning that at around 75 ms only 45% of the ASIS is still covered by the webbing. During the subsequent rebound phase, the coverage remains nearly constant. Fig. K2 represents a submarining case. Here, the z-distance becomes negative, resulting in a negative coverage value. This indicates that the ASIS is no longer covered by the lap belt, meaning that submarining has occurred.

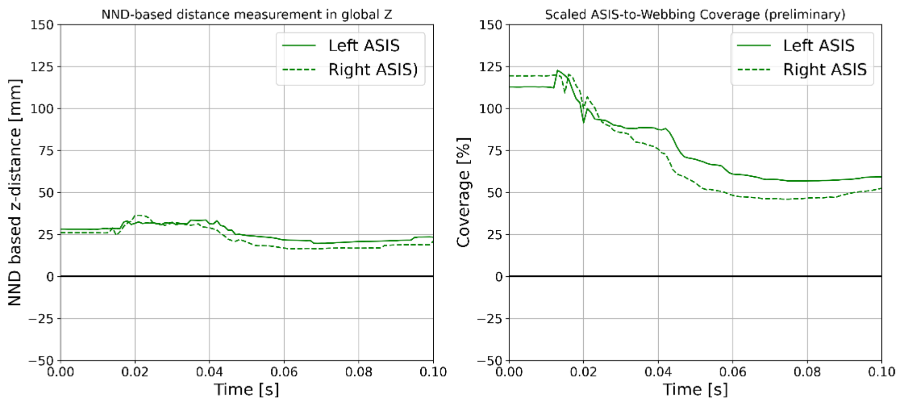

Fig. K1. Updated NND/NNP indicator for a non-submarining case.

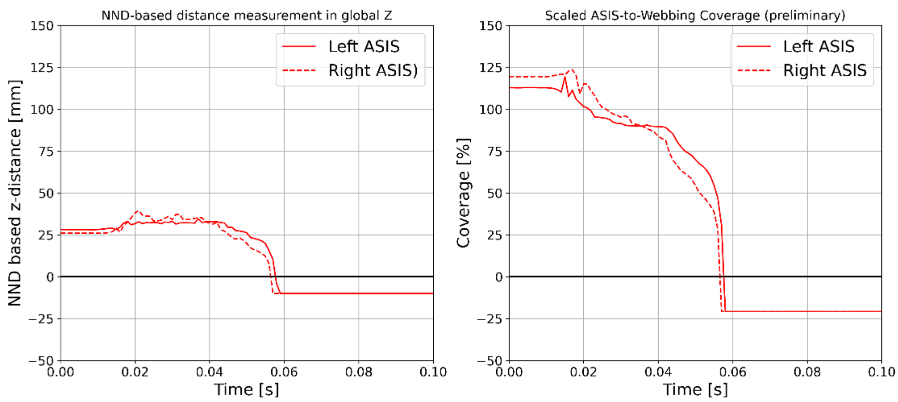

Fig. K2. Updated NND/NNP indicator for a submarining case.

Final comment on the terminology "Nearest Node":
Initially, the nearest node was chosen based on minimum distance in all directions. This proved error-prone due to belt bending and fluttering. Better results were obtained by selecting the nearest node in the y-direction but using the maximum distance in the x-direction relative to that node. Thus, the term "Nearest Node" may be misleading, but for consistency the original term is retained.

Appendix L
Force Drop evaluation

All Figures are related to HBM simulations in seat configuration [B2]. Figure Fig. L1 shows a simulation run with submarining occurrence only of the left ASIS side, no submarining on the right ASIS side. The force in the submarining case is slowly increasing over a long timeframe. Figures Fig. L2 and Fig. L3 show simulation runs with submarining on both sides (forces in kN, time in ms).

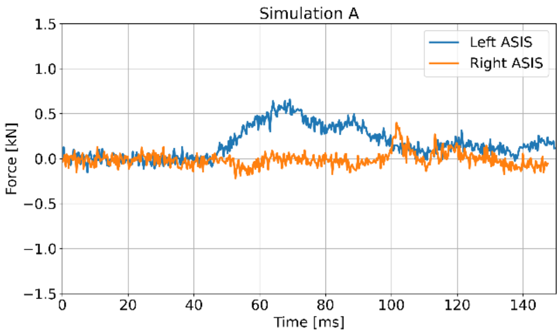

Fig. L1: Force Drop over Time, Simulation A.

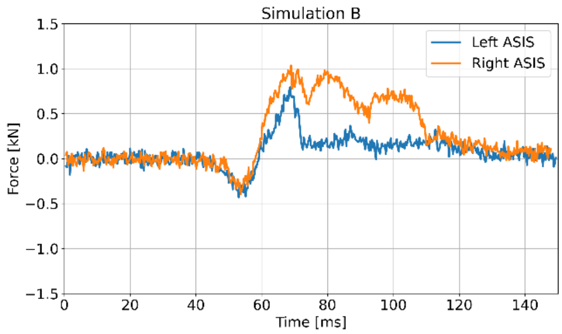

Fig. L2. Force Drop over Time, Simulation B.

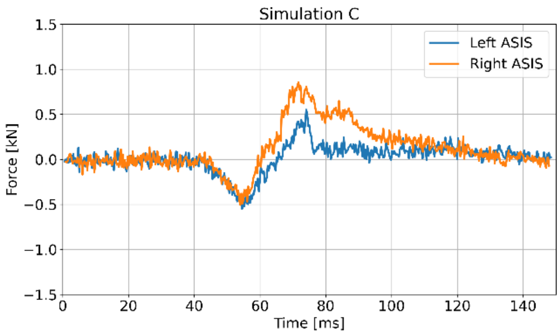

Fig. L3. Force Drop over Time, Simulation C.