Strain-based Rib Fracture Prediction Using HBMs: A Comparison of Simulations Using State-of-the-art Models to PMHS Tests Considering Uncertainty

Felix Ressi, Vishnuvel Kannan, Bernd Schneider, Corina Klug

Abstract Human body models (HBMs) are increasingly used to assess thoracic injury risk, yet model-to-model consistency remains uncertain. This study compared predictions of the risk of three or more fractured ribs (NFR3+) from nine Euro-NCAP-qualified HBMs across 17 thorax loading cases and related them to outcomes in post mortem human subjects (PMHS). Principal cortical rib strains were reduced with four scalar definitions and aggregated by 95th, 99th, and 100th percentiles, then mapped to rib fracture probability using a Weibull injury risk function (IRF) with age as a covariate and combined via a Poisson-binomial model. PMHS risk intervals were derived using Bayesian bootstrapping. Three calibration strategies for the strain to risk mapping were evaluated: count likelihood negative log likelihood, tail likelihood for NFR3+ negative log likelihood, and Brier score minimisation. In extreme cases, strains differed by about an order of magnitude across models, and maximum or near maximum percentiles outperformed 95th percentile strains. Calibration generally improved agreement with PMHS intervals, but no single strategy dominated, and the age effect conflicted with some datasets. The calibrated IRFs presented in this study were intended as diagnostics of model differences, not as model specific functions. Standardised procedures with uncertainty handling are recommended for virtual testing.

Keywords Human body models, Rib fracture prediction, Strain-based injury prediction.

I. INTRODUCTION

Human body models (HBMs) have been a key tool for research and development in the area of road traffic and vehicle safety in recent years. While challenges remain, they promise to offer several benefits over anthropomorphic test devices (ATDs). One of the main advantages is that they enable tissue level injury risk assessment. Given the high relevance of thoracic injuries in motor vehicle crashes, in particular with respect to AIS3 injuries in mid- to high-severity crashes [1-3], it is not surprising that research on using HBMs for strain-based rib fracture prediction is plentiful [4-8].

While HBMs have been used in research and vehicle development for quite some time, their adoption in virtual testing is a recent phenomenon. Notably, the 2026 European New Car Assessment Program (Euro NCAP) frontal impact protocol requires vehicle manufacturers to submit results using a qualified average male HBM [9]. Specifically, as part of the “sled and virtual testing” assessment, results from HBM simulations seated in a frontal sled test model need to be submitted. Four different seat settings and respective HBM postures are required [10]. However, this first implementation serves monitoring purposes and does not assess injury risk using the HBMs. If no HBM results are provided, the maximum available score for virtual testing is halved [9]. For the 2029 protocol update, the inclusion of a rib fracture risk assessment is being investigated.

Several approaches to predict rib fractures with HBMs have been reported. One method involves implementing material failure into the simulation model directly. Using accumulated stress or strain, evaluated at element level, individual elements, which exceed a pre-defined threshold are “eroded”, i.e., deleted [11-14]. In essence, this approach attempts to replicate fractures deterministically, as they occur in-vivo. While the theoretical ability to replicate reduced rib cage stability as a consequence of an increasing number of rib fractures is sometimes reported as an advantage with this method [14], it has several drawbacks. First, biofidelic representation of cracks is not realistic, given the typical mesh sizes in cortical bone rib elements of state-of-the-art HBMs (2-7 mm) and the crack propagation speed in cortical bone (>300 mm/ms [15]), as elements cannot be deleted fast enough. Secondly, the threshold for element erosion needs to be determined *a priori* [14]. Hence, while rib fracture risk has been shown – although to varying degrees – to be dependent on age [4-7], this threshold can only represent

F. Ressi (+43 316 873 30363, felix.ressi@tugraz.at) is a Project Senior Scientist, V. Kannan a PhD student and Research Assistant, B. Schneider a Researcher and C. Klug is Associate Professor at the Vehicle Safety Institute at Graz University of Technology in Austria.

a specific age. This eliminates the option to evaluate the results of one simulation with respect to several age groups.

Another option is a probabilistic approach, aiming to determine the risk for a specific number of fractured ribs (NFR) based on strains observed in rib cortical bone elements. This approach does not address the biofidelity limitations of the deterministic approach, since without considering failure, the strains can reach nonphysical levels where even the ribs found in the most robust human would have long failed. Nevertheless, the method is widely used to estimate rib fracture risk considering the influence of age [14]. Originally published in 2012 [4], subsequent improvements by updating the stepped deterministic risk curve with a survival analysis-based smooth Weibull risk function [8] and the addition of an age-dependent rib strain-based fracture risk function built on new material data [5] were proposed. For two HBMs, model-tuned risk functions based on results of 170 simulations covering 13 impactor and tabletop load cases are available [6].

In the 2026 monitoring phase, Euro NCAP has opted to allow any HBM to be used that fulfils the qualification procedure outlined in Technical Bulletin (TB) Crash Protection (CP) 550 [16] rather than only allowing just one specific HBM. However, previous research suggests using the age-dependent rib strain-based fracture risk function built on new material data [5] directly can result in very different rib fracture predictions from state-of-the-art HBMs under identical loading conditions [6][17]. Hence, to ensure comparable results with respect to rib fracture risk, it seems model-specific rib fracture risk curves would be needed for each qualified model. While feasible, this would increase intrinsic uncertainty, because (i) adopting the approach presented in [6] as an example, the curves would be tuned on a limited set of 170 simulations and may not generalise to other load cases, postures, or restraints; (ii) tuning may mask model-specific differences in strain scale; (iii) each model update would necessitate a new, version-specific curve with unknown behaviour outside the calibration set; and (iv) the small and variable PMHS samples underlying the fits yield wide statistical uncertainty that propagates differently across models.

In view of these limitations, the study is presented as an evaluation with three key objectives:

- To compare HBM predicted NFR3+ risk with PMHS based statistics, with uncertainty represented.
- To calibrate, post hoc, the strain to risk mapping applied to analysed rib cortical strains. Note the resulting injury risk functions (IRFs) are not intended to be used as model-specific functions but as diagnostics of model differences.
- To compare NFR3+ risk estimates obtained from the original material-based mapping, from the calibrated mapping, and from published model specific curves where available.

II. METHODS

The present study is based on two essential elements: statistical analyses of existing PMHS data and strain-based rib fracture risk assessment of HBMs.

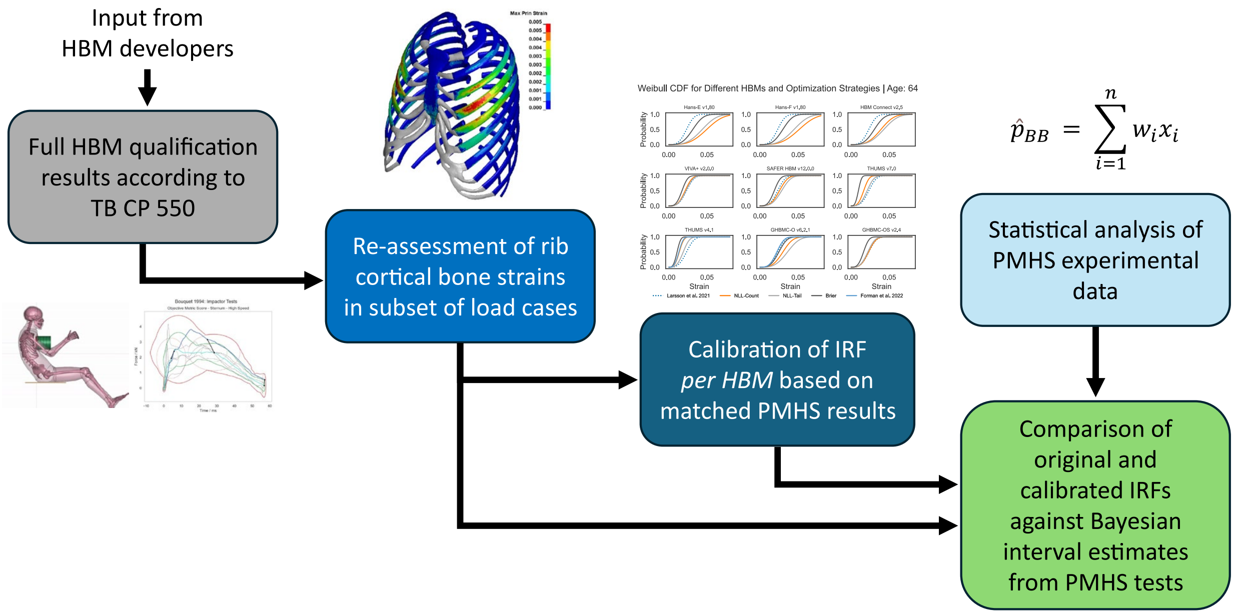

Fig. 1. Flowchart of the approach applied in the present study.

The flowchart in Fig. 1 illustrates how the two were combined to compare rib fractures observed in tests with strain-based rib fracture risk predictions and calibrate injury risk function shape parameters to highlight the differences between the HBMs.

As the starting point for the rib fracture assessment, a subset of the data submitted by the model developers for HBM qualification for Euro NCAP [16] was analysed. Specifically, the strains in the cortical bone parts of the ribs were assessed. These were then used directly as input to a probabilistic rib fracture risk prediction method from literature [6]. To highlight the differences between the HBMs' strain responses under identical boundary conditions further, calibration strategies were applied, minimising the difference between predicted and observed injury risk in paired PMHS tests for each HBM individually. The final comparison with the PMHS data involved Bayesian bootstrapping to better capture and quantify the underlying variability.

The following subsections explain the approach in more detail.

Statistical Analysis of PMHS Results

As the present study sought to utilise rib strain data already submitted as part of the full HBM qualifications according to TB CP 550 [16], load cases were selected from the HBM4VT validation catalogue [18-19]. The following inclusion criteria were used to choose 17 out of the 50 load cases (which are grouped into 19 validation setups):

- Loads applied to the rib cage
- Data from at least three PMHS available
- Minimum score thresholds defined in TB CP 550 [16]
- Post-test injury information available
- Inclusion of male subjects only

The rationale behind these selection criteria is detailed in the following. The selection was limited to cases, where loading was applied to the thorax, i.e., body region of interest in the present study. A minimum sample size of three was defined in an effort to increase trust in test reproducibility and to reduce the effect of potential outliers. Furthermore, only load cases where a minimum threshold for at least one of the two objective metric scores applied in TB CP 550 [16] was defined. The motivation for this criterion was that this provided a minimum level of trust in the simulation setup's capability to replicate the respective physical experiments. In addition, only load cases were included, where enough post-test injury information was available to determine the number of fractured ribs for each test. Finally, for cases where female and male PMHS were tested, the selection was limited to male subjects.

TABLE I

OVERVIEW OF UTILISED LOAD CASES

Validation setup	Load case	Reference	PMHS ages	# NFR3+
Bouquet *et al.* 1994	Sternum high speed	[20]	76, 57, 66, 69	3
Kroell *et al*. 1971	High speed	[21]	19, 29, 72, 78, 80, 81	4
Lopez Valdes *et al*. 2018	Restraint condition a	[22]	68, 74, 94	3
	Restraint condition b	[22]	63, 73, 74	1
Petit *et al*. 2019	Far-side Sled	[23]	74, 82, 88, 73, 87, 91	4
Salzar *et al*. 2009	Load Condition 1	[24]	62, 54, 31	0
	Load Condition 2	[24]	62, 54, 31	0
Shaw *et al*. 2009	Gold Standard 2	[25]	59, 66, 67, 67, 74	1
Uriot *et al*. 2015	Seat Configuration front	[26]	85, 84, 84, 89	4
	Seat Configuration rear	[26]	55, 86, 87, 87	4
Viano 1989	Abdomen low speed	[27]	29, 62, 64	0
	Abdomen medium speed	[27]	38, 66, 64	2
	Abdomen high speed	[27]	49, 29, 62, 64	2
	Thorax low speed	[27]	29, 37, 64	0
	Thorax medium speed	[27]	38, 64, 66, 63	4
	Thorax high speed	[27]	49, 49, 29, 64, 37	5
Yoganandan *et. al* 1997	Hub	[28]	72, 81, 84, 86, 62, 70, 68	5

Table I above lists all selected load cases with the ages of the tested PMHS and the resulting number of cases where the number of fracture ribs was three or higher (NFR3+). Note that four of the nine validation setups in this list were also utilised to tune model-specific injury risk functions for THUMS v4.1 and GHBMC v6.0 [6], specifically Kroell *et al.* 1971 [21], Salzar *et al.* 2009 [24], Shaw *et al*. 2009, and Yoganandan *et. al* 1997. However, based on the filter criteria listed above (the omission of female data in particular), only a subset of PMHS from these publications were considered in the present study.

Due to the limited sample size of PMHS tests and the inherent variability between subjects, direct comparison of HBM predicted NFR3+ risk to PMHS injury outcome can be highly uncertain and sensitive to individual observations. To address this, Bayesian bootstrapping (BB) was applied to estimate the mean NFR3+ risk and its associated uncertainty through interval estimation.

The PMHS NFR3+ data were grouped by loading condition. For each loading condition with n subjects, the NFR outcomes for each subject i were treated as binary variables $x_i \in \{0, 1\}$, where 1 indicates NFR3+ occurrence.

Bayesian bootstrapping [29] was performed by assigning random Dirichlet weights, $w \sim Dirichlet\ (1, \dots ,1)$ to the NFR 3+ outcomes. Then the weighted estimate $\hat{p}_{BB}$ of NFR3+ was obtained by the formula provided in Equation (1) below:

$$\hat{p}_{BB} = \sum_{i=1}^{n} w_i x_i \quad (1)$$

This process was repeated over many iterations to generate a posterior distribution of weighted risk estimates, from which 95% highest density intervals (HDIs) were derived. This approach translates the limited and variable PMHS samples into a 95% HDI for NFR3+ risk. An HBM prediction that falls within this interval can be regarded as consistent with the available PMHS evidence, given the uncertainty.

Strain-based Rib Fracture Risk Assessment of HBM Simulations

To qualify an HBM for application according to Euro NCAP's Frontal Impact Protocol [9] in simulations described in TB CP 551 [10], all requirements outlined in TB CP 550 [16] must be satisfied. Part of these requirements is the submission of simulation results to Euro NCAP that demonstrate the HBM's biofidelity. In every evaluation described for these simulations, the minimum objective metric score must be met [16][19]. In the present study, the binary simulation results of nine HBMs were analysed following their qualification. TABLE II lists the considered HBMs, their version numbers and the unit systems they were qualified in.

TABLE II
OVERVIEW considered HBMs

Name	Version	Unit system
Hans E	v1.80	kg-mm-ms
Hans F	v1.80	kg-mm-ms
HBM Connect	v2.5	kg-mm-ms
VIVA+	v2.0.1	kg-mm-ms
SAFER HBM	v12.0.0	kg-mm-ms
THUMS	v4.1	t-mm-s
THUMS	v7.0	t-mm-s
GHBMC-O	v6.2.1	kg-mm-ms
GHBMC-OS	v2.4	kg-mm-ms

All simulations had been performed by the HBM developers using the explicit finite element (FE) solver LS-Dyna (Ansys, Inc. Canonsburg, PA, USA). Only results from the 17 load cases listed in Table I were considered.

For these simulation results, rib fracture risks were assessed using the open-source Python library Dynasaur+ v1.1.0 [30] using the procedure outlined below:

For each shell element of the rib cortical bone and each time step, the 3×3 symmetric strain tensors at the upper and lower through-thickness integration points were assembled from LS-Dyna outputs. Principal strains were then reduced to a single scalar per element using four variants that differ by the order of through-thickness combination and by how compressive and tensile strain components are treated:

- mean-overall: First compute the mean strain tensor as the element-wise average of the upper and lower tensors. Then compute its principal strains and report the largest absolute principal strain.
- mean-tension: First compute the mean strain tensor as above. Then compute its principal strains and report the largest nonnegative principal strain (i.e., the maximum of the three principal strains and zero), thereby ignoring compression.
- overall-max: Compute principal strains separately at the upper and lower integration points. Report the maximum absolute value across all six principal strains.
- tension-max: Compute principal strains separately at the upper and lower integration points. Report the maximum of those six principal strains and zero, thereby ignoring compression.

These four principal strain scalars were then aggregated at the rib level by computing, for each time step, the 95th, 99th, and 100th (maximum principal strain, MPS) percentiles across elements, yielding strain-time histories for each rib. This investigation of different percentiles was included to investigate whether feeding different strain percentiles into the rib fracture risk functions would result in improved prediction accuracy, loosely following an approach previously considered for brain injury prediction [31].

The peak strain from each rib's time history was then used to calculate the fracture probability for each rib, using the Weibull cumulative distribution function (CDF), which uses age as a covariate [6]. The individual per-rib fracture ribs were then combined into an overall NFR3+ risk with a Poisson-binomial count model [4][6].

In an effort to further illustrate the differences in strain response from the nine HBMs considered in this study, similar to the approach outlined in [6], the rib fracture risk function for each HBM was calibrated to minimise prediction error between the strain-based NFR3+ risk and the injury outcome in the matched PMHS tests.

Specifically, three different calibration strategies were considered for estimating the Weibull scale parameters (β_0, β_1). In all cases, per-rib fracture probabilities were computed from simulated rib strains and subject age via a Weibull CDF with fixed shape α, consistent with the approach presented by Forman *et al.* [6].

1. Count likelihood ("NLL-Count")
   Parameters were estimated by maximising the Poisson-binomial likelihood of the observed rib-fracture counts per test, equivalently minimising the average negative log-likelihood (NLL) of the counts. This approach uses the full count information.
2. Tail likelihood for NFR3+ ("NLL-Tail")
   For the indicator of "3 or more fractured ribs" (NFR3+), parameters were estimated by maximising the Bernoulli likelihood of this indicator, using the Poisson-binomial tail probability provided in Equation (2) as the predicted probability (i.e., minimising log-loss/cross-entropy).

$$S = P(N \geq 3) \tag{2}$$

   This strategy targets the relevant threshold directly. For binary classification log-loss is defined in Equation (3), where y_i are the true labels and p_i the predicted probabilities.

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{3}$$

3. Brier-score minimisation for NFR3+ ("Brier")
   Parameters were estimated by minimising the average Brier score for the NFR3+ indicator, using the tail probability provided in Equation (2). This emphasises calibration in squared-error terms and is less sensitive to extreme misclassifications than log-loss. Brier-score is defined in Equation (4), where y_i are the true labels and p_i the predicted probabilities.

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \tag{4}$$

All three strategies were repeated for the four principal strain scalars mean-overall, mean-tension, overall-max, and tension-max. The IRFs resulting from these three calibration approaches were then compared with the results from the original function and with model-specific risk curves where available.

III. RESULTS

To illustrate the effects of different percentile strains and the four principal strain scalars, Fig. 2 shows boxplots of the maximum strains observed in all 24 ribs for each HBM for the load case Kroell *et al.* 1971 high speed [21]. Equivalent plots for the other 16 load cases are provided in the Appendix.

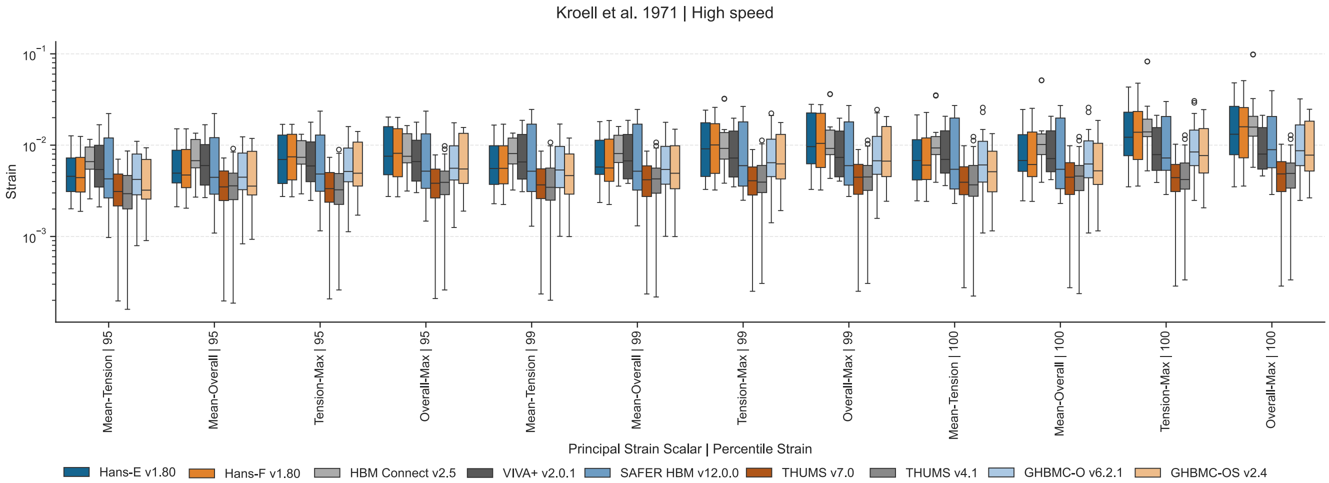

Fig. 2. Boxplots of maximum strains for all 24 ribs for each HBM in one load case [21]. Note that a logarithmic axis was used for the strains.

It is evident from Fig. 2 that the strains increase gradually with increasing percentiles, as is to be expected. Furthermore, within each percentile group, the maximum strain values calculated for the four principal strain scalars follow a consistent order. Specifically, mean-tension yields the lowest values, followed by mean-overall, tension-max and overall-max.

The upper plot in Fig. 3 compares the NFR3+ risks from all HBMs obtained by using the original IRF presented in [5] using the principal strain scalar tension-max. In the background, Bayesian bootstrapping interval estimates from the matched PMHS tests are indicated.

In several load cases in Fig. 3, no interval estimates are provided. This is because in 9 out of the 17 load cases, the Bayesian bootstrapping interval estimates collapse, reflecting all-or-nothing outcomes among the PMHS: NFR3+ in all subjects (Lopez Valdes *et al*. 2018 RCa, Uriot *et al*. 2015 SC front and rear, Viano 1989 Thorax high-speed and medium-speed) or in none (Salzar et al. 2009 LC1 and LC2, Viano 1989 Abdomen low-speed and Thorax low-speed). Nevertheless, to clearly indicate whether three or more fractures were observed for *all* or *none* of the PMHS in these cases, a thin grey bar was added to the corresponding side of the plot indicating an estimate of 100% or 0% respectively.

When comparing the HBM results from the original IRF with the interval estimates from the PMHS tests, it is evident that in the two Uriot *et al.* 2015 load cases, they agree perfectly, with all HBMs predicting 100% NFR3+ risk. There are three further load cases, where the majority of HBMs aligns well with the interval estimates from the PMHS data. Specifically, Salzar *et al.* 2009 LC1 and LC2, and Viano 1989 Thorax high speed. While in the former, none of PMHS exhibited three or more fractured ribs, in the latter, NFR3+ occurred in all tests. This is predicted well by the HBMs, with maximum NFR3+ predictions of around 10% and 20% in Salzar *et al.* 2009 LC1 and LC2 respectively and the minimum prediction of approx. 82% in Viano 1989 Thorax high speed.

Apart from these extreme cases, no overall trend is evident from a load case perspective. It is noteworthy though, that some HBMs tended towards overpredictions while others tended towards underpredictions.

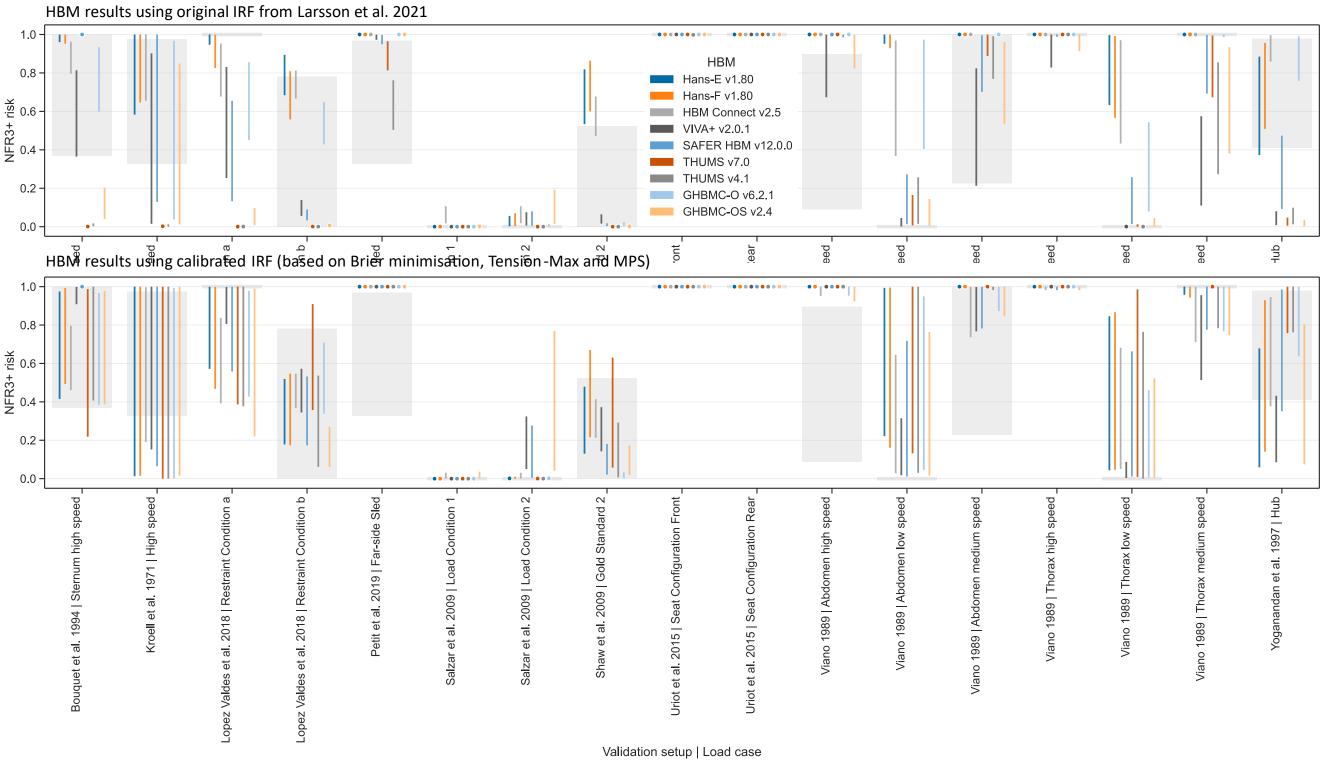

Fig. 3. Line plots showing the range of NFR3+ risk for each HBM by validation setup and load case. The upper plot shows results using the original material-based IRF published by Larsson *et al.* [5]. The lower plot shows results using the calibrated HBM-specific IRFs based on Brier minimisation and the MPS from the tension-max principal strain scalar. For both, the age covariate in the NFR3+ calculation was set to the age of the individual PMHS, resulting in the range of risk values. Background bands show interval estimates from matched PMHS tests using Bayesian bootstrapping (light grey).

Following calibration, the results were analysed in the same way. The resulting line plots are presented in the lower plot in Fig. 3. Comparing the upper and lower line plots, the HBMs' NFR3+ risk predictions are generally in better agreement with the PMHS tests for most load cases, clearly illustrating the desired effect of calibration. The clearest examples for this are Bouquet *et al.* 1994, Lopez Valdes *et al.* 2018 RCb, and Shaw *et al.* 2009 Gold Standard 2. Prior to calibration, there were several HBMs which either under or overpredicted the injury risk when compared to the interval estimates. After calibration, the line plots for all HBMs fall within the interval estimates of these three load cases, apart from small individual deviations. Hence, they can be considered consistent with the available PMHS evidence. In many other load cases, while the improvement may be less clear or not as uniform for all HBMs, the overall alignment is at least improved. Examples of this group are Lopez Valdes *et al.* 2018 RCa, Salzar *et al.* 2009 LC1, Viano 1989 Thorax high speed, Thorax medium speed, and Yoganandan *et al.* 1997. For Kroell *et al.* 1971, Viano 1989 Abdomen low speed and Thorax low speed, the effects are mixed, with some HBMs improving their alignment and others reducing it due to much wider ranges of results. For both load cases in the Uriot *et al.* 2015 validation setup, no change was observed, with all HBMs still correctly predicting 100% NFR3+ risk. Finally, for Petit *et al.* 2019, Salzar *et al.* 2009 LC2, Viano 1989 Abdomen high speed, and Viano 1989 Abdomen medium speed, the calibration resulted in reduced agreement between the HBMs' predictions and the PMHS interval estimates. While it may seem counter intuitive that the results deteriorate following calibration, this can be explained by the nature of the calibration process. Specifically, each individual PMHS result is considered equally important. Reduced scores or reduced PMHS-HBM NFR3+ alignment in individual load cases may occur, as long as this is compensated by a larger increase in other instances as this ensures that the overall result (i.e., reduced Brier score in the example provided in Fig. 3) is still improved.

To investigate potential reasons for the different degrees to which calibration was able to improve alignment between PMHS tests and HBM NFR3+ predictions, plots showing the number of fractured ribs in PMHS tests as function of subject age were created. In addition, mean HBM NFR3+ risk predictions and respective interquartile ranges (IQRs) were added for the original and calibrated IRF (based on Brier approach). Fig. 4 shows such plots for two load cases, Bouquet *et al.* 1994 and Petit *et al.* 2019. Notably, the HBM-specific IRF calibration, which effects on the individual load case results are illustrated in Fig. 3, led to improved agreement between PMHS results and the prediction from most HBMs for Bouquet *et al.* 1994. In contrast, the calibration did not improve HBM predictions for Petit *et al.* 2019.

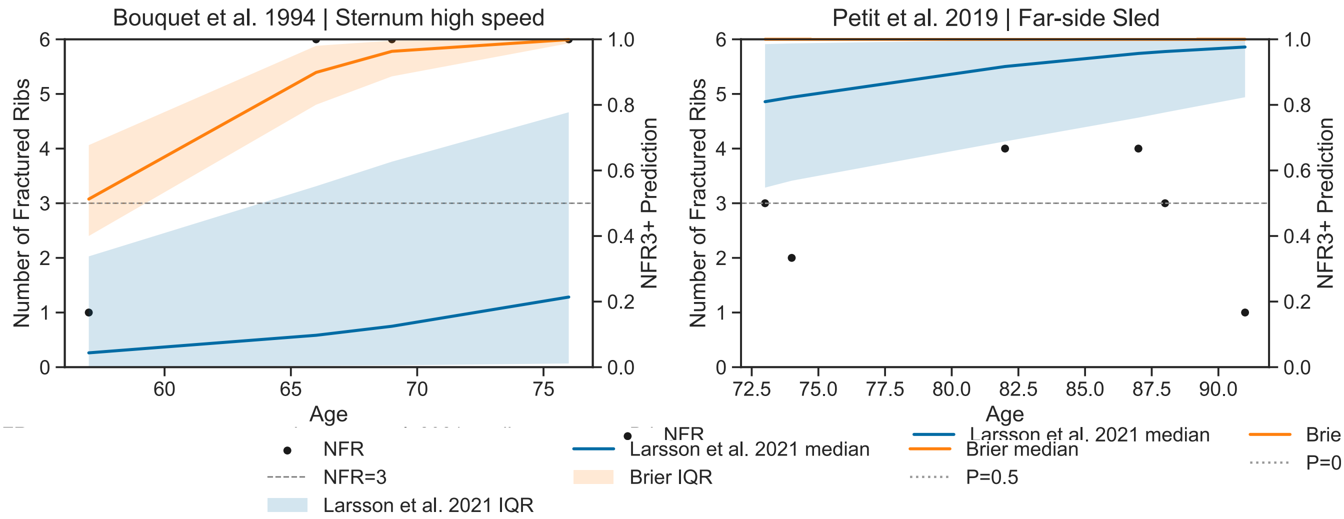

Fig. 4. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots). Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+. Exemplary plots shown for Bouquet *et al.* 1994 Sternum high speed [20] (left) and Petit *et al.* 2019 [23] (right). Equivalent plots for all load cases are provided in the Appendix.

When comparing the two plots side by side, it is obvious that while for Bouquet *et al.* 1994 the number of fractured ribs increases with increasing age, this is not the case for Petit *et al.* 2019. Since the commonly used method applied for rib fracture prediction uses age as a covariate in the IRF, fracture risk increases monotonically with increasing age. This is in line with the PMHS data from Bouquet *et al.* 1994 with lower NFR for the youngest subject. However, this assumption is not valid for the PMHS tests presented in Petit *et al.* 2019, as there is no evidence for a monotonic relationship between age and NFR, as illustrated by the right-hand side plot in Fig. 4.

While Fig. 3 provides insights into the NFR3+ response of all HBMs in each load case before and after calibration for one principal strain scalar (tension-max), strain percentile (100th) and calibration strategy (Brier), it is impractical to use this type of visualisation to compare the three calibration strategies and investigate the effects of the four principal strain scalars.

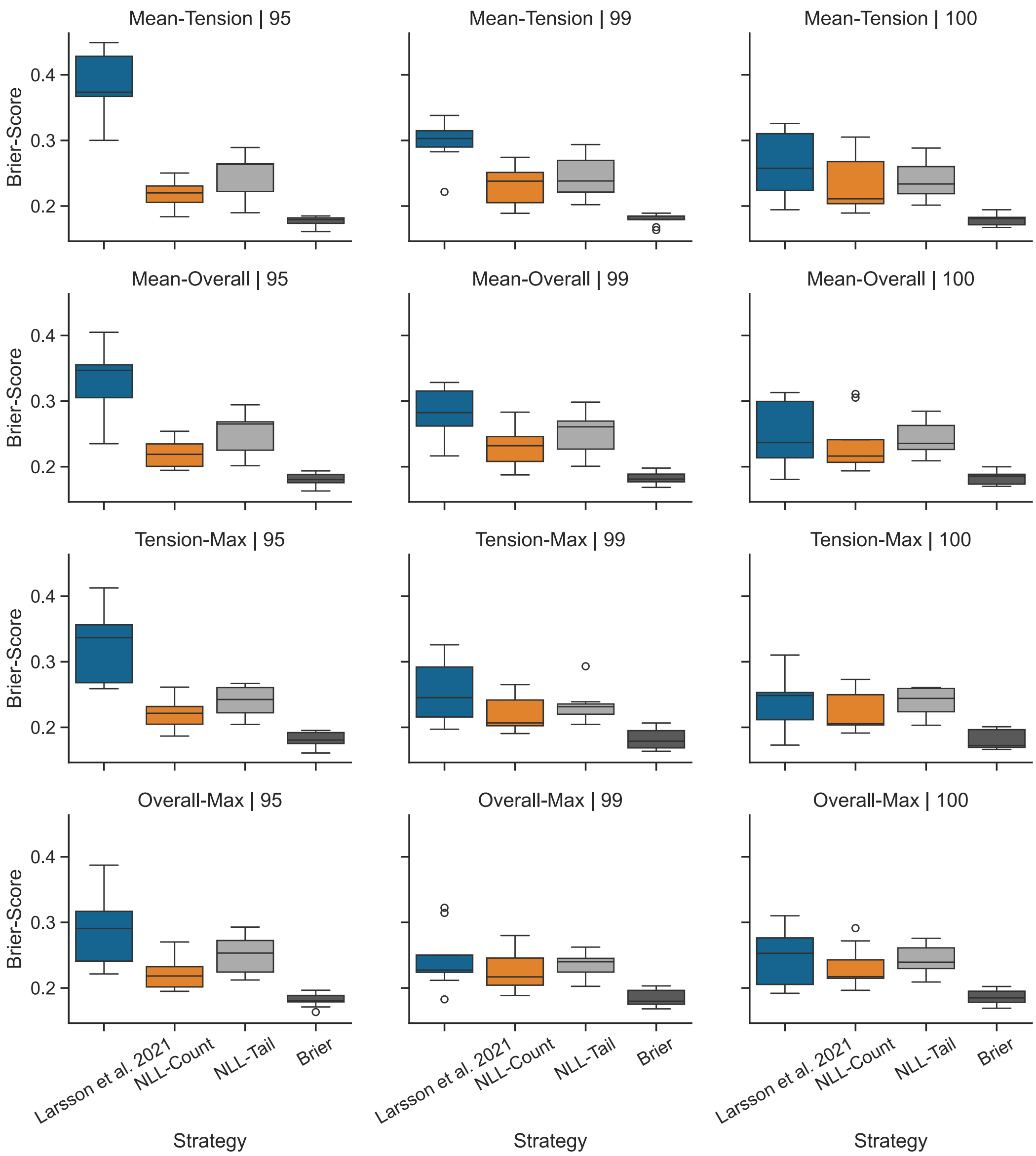

Fig. 5. Comparison of calibration results for the three strategies and the baseline (Larsson *at al.* 2021 [5]) three. Brier scores (lower is better) are provided, regardless whether it was used as calibration target. The 12 plots show all combinations of strain percentiles and principal strain scalar options.

For this purpose, the minimised strategy targets (log-loss or Brier values) were analysed for all 12 combinations of principal strain scalar and strain percentile considering the three calibration strategies. While detailed bar charts showing the results per HBM are provided in the Appendix, results of all HBMs aggregated into box plots were created for easier comparison. Fig. 5 shows the respective results using Brier scores on the y-axes. The equivalent plot using log-loss is provided in the Appendix.

Analysing the blue box plots for the original IRF [5] in the 12 subplots, it is evident that the Brier scores reduce (i.e., improve) for the considered principal strain scalars, with mean-tension yielding the highest values, followed by mean-overall, tension-max and overall-max (from top to bottom in Fig. 5). Median Brier scores also consistently reduce with increasing percentiles for the mean-tension and mean-overall principal strain scalars, with the 95th percentiles resulting in the highest Brier scores, followed by the 99th and the 100th (i.e., the maximum principal strains). For the tension-max and overall-max principal strain scalars, the lowest Brier scores were obtained with the 99th percentile strains.

Brier scores for the log-loss-based calibration strategies were consistently higher compared to the strategy based on Brier score. As is to be expected, the opposite is true when plotting the results with log-loss on the y-axes (see Fig. A17). Notably, while the spread in Brier score results across HBMs is considerably smaller in certain principal strain scalar and strain percentile combinations, the overall lowest median Brier score of approximately

0.17 is achieved for tension-max combined with the 100th percentile strains. This illustrates that the choice of principal strain scalar can affect the resulting NFR3+ risk prediction just as much as the choice of strain percentile.

In order to illustrate the differences between the calibrated IRF for each HBM, the resulting Weibull CDFs of single rib fracture risk as a function of strain were plotted. Fig. 6. shows these Weibull CDFs for all HBMs evaluated at an age of 64 years, matching the mean PMHS age used in this study. This is also consistent with the mean PMHS age reported in [6]. Calibrated IRFs from NLL-Count, NLL-Tail, and Brier are compared with the baseline IRF from [5]. For THUMS v4.1 and GHBMC-O v6.2.1 the HBM-specific IRFs from [6] are also included. Across most HBMs, the NLL-Count and NLL-Tail calibrations align closely, whereas the Brier-based IRF is shifted towards lower strain values. Relative to the baseline IRF, the calibrated curves for Hans-E v1.80, Hans-F v1.80, and HBM Connect v2.5 are shifted towards higher strain values, indicating overprediction in these models when using the baseline IRF. In contrast, for both THUMS models the calibrated IRFs lie to the left of the baseline, indicating underprediction with the baseline IRF. Notably, for THUMS v4.1 the NLL-Count calibration is closest to the model-specific IRF reported in [6], while for GHBMC-O v6.2.1 the Brier-based IRF is closest. For VIVA+ v2.0.1, SAFER HBM v12.0.0, and GHBMC-OS v2.4, the calibrated IRFs remain near the baseline IRF, suggesting comparatively low sensitivity to the choice of calibration approach.

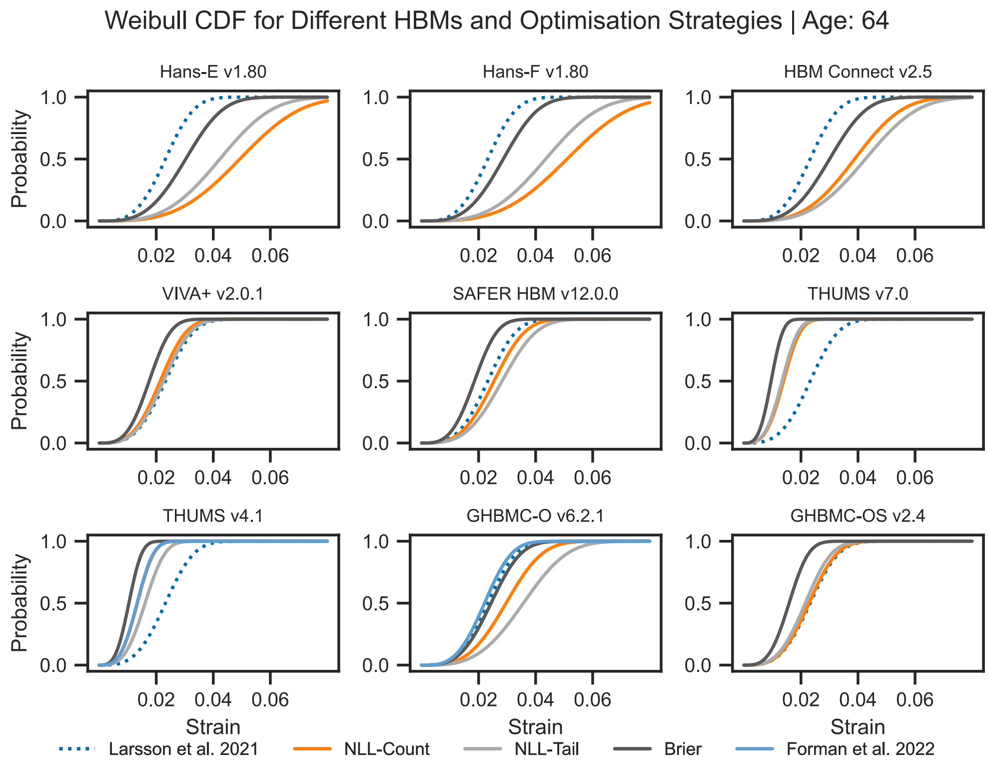

Fig. 6. Weibull CDFs of single rib fracture risk as a function of strain for all HBMs for an age of 64. Each subplot overlays the IRFs fitted via different optimisation strategies (NLL-Count, NLL-Tail, and Brier) with the baseline IRF from [5]. Model-specific Weibull CDFs are added from [6] where applicable.

Because the shape and position of the IRFs are highly age-sensitive, calibrated IRFs were also evaluated at age 45 for all HBMs (Fig. A35), enabling a direct comparison with the plots in [6].

The calibrated Weibull parameters are not reported directly, as the intent is to use the resulting IRFs solely as diagnostics of model differences rather than as model-specific functions for rib fracture prediction.

IV. DISCUSSION

One goal of this study was to robustly compare HBM-predicted NFR3+ risk to PMHS outcomes in a way that reflects uncertainty. Because PMHS sample sizes are small and vary across subjects, any single estimate can be driven by a few tests. To somewhat overcome this issue, Bayesian bootstrapping was used to estimate the interval of the PMHS NFR3+ risk. The Bayesian bootstrap summarises only what is in the observed PMHS data by resampling and reweighting; it is a faithful finite-sample summary, but when all outcomes are identical (all zeros or all ones) its interval collapses to exactly 0% or 100% with no width, which can give a false sense of certainty. However, this also makes sense intuitively, as there simply is no data to motivate any interval width in this scenario. One key shortcoming with respect to Bayesian bootstrapping in the present application is that the age information from the PMHS is not considered in the resampling. Given the well-established age-dependency of rib fracture risk [4-7], this represents an important limitation. Future research could investigate ways to improve the presented approach to account for age. Nevertheless, given the small sample size typically associated with PMHS tests, providing interval estimates for PMHS results seems reasonable in general as it visualises the inherent uncertainty related to human variability.

While small sample size and age distributions skewed towards older subjects are common issues in PMHS tests used for rib fracture prediction [32], another limitation of the experimental data considered in this study is the outcome distribution. Specifically, in four of the 17 load cases none of the PMHS exhibited three or more fractured ribs, whereas in five load cases all tests exceeded this threshold. While such all-or-nothing cases can still contribute to validating tissue-based predictions, it would be preferable to include load cases that yield a spectrum of outcomes, particularly around the NFR3+ threshold, as these carry the most information for estimating and validating risk functions. In one of the no-fracture load cases, the same subjects were tested twice [24]. Given the uniformly non-injurious outcomes, this duplication is unlikely to bias the analysis, though it slightly reduces the effective sample size. In addition, some PMHS were tested twice on opposite sides of the torso in two impactor load cases (Viano 1989 Abdomen low speed and Thorax low speed) [27]. To preserve independence, only the first trial per subject was used.

Comparing the direct strain outputs from nine HBMs revealed considerable differences between them, even though the results were obtained from simulations under well-controlled boundary conditions. Analysis of the strain outputs showed that the highest strain values for some of the nine HBMs, particularly when analysing the 100th percentile of the overall-max principal strain scalar, can be an order of magnitude apart. Given this, applying one IRF to all models considered in this study is bound to result in considerable differences in predicted NRF3+ risk. While one of the qualified HBMs was validated to reproduce rib strain measurements from PMHS tests [7][33], to the best of the authors' knowledge, similar validation for the remaining models has not been reported.

The relatively small differences when investigating maximum strain values across the two tension- and overall-pairs (i.e., mean-tension compared to mean-overall and tension-max compared to overall-max), are in line with findings from a study in which strain gauges were applied to the ribs of PMHS subjected to belt loading in a table top test. There, the strain gauge data confirmed that tensile loading was the predominant mode, most relevant for rib fracture [34].

However, even with these inherently different strain outputs from the HBMs, several load cases were identified, in which all HBMs overpredicted NFR3+ risk in practically all simulations. For these load cases, this points less towards an issue in the HBMs' modelling approaches and more to a common difference between the PMHS tests and the simulation setups [19]. At the time of preparing this manuscript, updates to the HBM4VT validation catalogue were in development. Some of these could improve the alignment between PMHS tests and the simulation model, thus improving the HBM response.

Moreover, the calibrated IRFs were broadly consistent with published model-specific curves, with different optimisation strategies aligning best for the two HBMs where model-specific IRFs were available, supporting their use as diagnostics of model differences. While some over and underpredictions were still observed after the IRF calibration, the results were more comparable among the models. While highlighting the power of model-specific calibration, this also illustrates an underlying problem: Tuning model-specific IRFs can mask underlying differences, e.g. in strain scale. Simply changing the strain to risk mapping does not improve the underlying difference in biofidelity.

Based on the direct strain results presented in Fig. 2 and the Appendix, omitting even just the highest 1% of strains likely discards too much data needed for accurate rib fracture prediction in some cases. Notably, to the best knowledge of the authors, the model-specific rib fracture functions presented in [6] rely on MPS derived from the mean of the through-thickness integration points, considering tensile and compressive strains, even though the original publication states that the 95th percentile strains were used. Consistent with that implementation, the results of the present study indicate that using 100th percentile strains with tensile-based strain scalars, or 99th percentile strains with principal strain scalars considering overall strains, yields more accurate predictions than 95th percentile strains.

Three calibration approaches were described. While not the exact tuning framework was not explained in detail in [6], the NLL Tail approach was modelled to replicate the tuning framework used in that study. As expected, NLL Tail calibration generally yields the lowest log loss for NFR3+, because it directly minimises cross entropy for that binary target. Likewise, Brier score minimisation tends to yield the lowest Brier score. Log-loss is unbounded and strongly penalises “confidently wrong predictions” (probabilities near 0 or 1 assigned to the opposite outcome). The Brier score is bounded in [0, 1] and therefore less sensitive to such extremes. In some load cases, older subjects exhibited fewer fractures, even at the NFR3+ threshold (as illustrated in Fig. 4), contrary to the model’s monotone age effect. Under this misspecification, Brier score minimisation can be preferable for NFR3+ because it is more robust to a small number of discordant, high penalty cases. However, this comes with trade-offs: it does not optimise log-loss and may yield less sharply informative probabilities. NLL Count optimisation can be preferable when a single model is required to provide calibrated probabilities across multiple thresholds K (NFRK+) or for the full count distribution. By fitting the Poisson-binomial likelihood to exact counts, it uses all available information and tends to transfer well across thresholds. However, because the probabilistic framework does not model element failure or loss of rib cage integrity, strains can grow to non-physical magnitudes and inter-rib dependence is not captured. Consequently, assessing risks for higher thresholds (beyond NFR3+) is not advisable, which limits the practical value of NFRK+ approaches.

Limitations

There are several limitations with respect to the present study. A key difference between the used qualification setups and those used for creating the model-specific risk function presented in [6] is their intent. While the models set up to obtain the model-specific IRFs were carefully developed to replicate PMHS position in each load case, the simulation setups available in the HBM4VT validation catalogue are designed to be HBM-neutral [19]. Hence, for instance, the HBM position is not perfectly matched to the PMHS position, but rather the setup was designed in a way to ensure reasonable reproducibility. This approach enhances comparability between HBMs, a key requirement for the analyses reported here.

As highlighted above, the setups in the HBM4VT validation catalogue are currently being updated. As soon as new HBM qualification results from these updated setups are available, the approach presented in this study should be reassessed.

Due to limited availability of qualification results from other FE solvers, such as Structural Mechanics VPS (Keysight Technologies Inc., Santa Rosa, CA, USA), at the time of preparing this study, only results using LS-Dyna were assessed. Future work should investigate the transferability of the IRFs between solvers.

As the intent of the calibrated risk curves was to highlight the differences between the individual HBM responses, no attempt to verify their validity was included in the present study. Prior to the adoption of new risk curves, studies investigating stochastic replication of real-world crashes [6][17] could be used to assess their robustness.

V. CONCLUSION

To the best knowledge of the authors, this is the first study comparing strain-based rib fracture risks across nine state-of-the-art human body models using 17 standardised simulation setups. By explicitly quantifying experimental uncertainty in PMHS outcomes using Bayesian bootstrapping, the evidential value of limited datasets was increased. Large differences in rib strain magnitude were observed across HBMs under identical boundary conditions, with extreme cases approximately an order of magnitude apart. These differences alone explain substantial divergence in NFR3+ risk when a single material-based IRF is applied. Maximum or near-maximum percentiles (99^{th}-100^{th}) outperformed 95^{th} percentile strains, with small differences between 99^{th} and 100^{th} depending on the principal strain scalar. Post hoc calibration of the strain to risk mapping improved agreement with PMHS-based interval estimates for most load cases. The calibrated IRFs should be interpreted as diagnostics of model differences, not as model specific IRFs for rib fracture prediction. No single calibration strategy achieved uniformly good agreement across all load cases. This reflects limitations of an age only covariate in the IRF and heterogeneity in PMHS outcomes, where fracture risk does not always increase monotonically with age.

For future use in virtual testing, confidence in HBM based rib fracture predictions should be ensured through a clear, standardised methodology. The assessment procedure should be standardised, reproducible and transparent, and validation against relevant evidence (rib fracture outcomes and, where feasible, rib strain measurements) should be part of the models’ qualification, with uncertainty handled explicitly. This would support consistent interpretation of results across models and future model updates.

VI. ACKNOWLEDGEMENTS

The authors acknowledge the IT Services (ZID) of Graz University of Technology, Austria, for providing access to high-performance computing resources. We further thank all HBM developers for submitting the results of their full HBM qualifications and for making the data available for the research presented in this study. This document reflects only the authors' view; the developers of the HBMs considered in this study are not responsible in any way for any use that may be made of the information it contains.

AI Disclosure: Generative artificial intelligence tools were used in context with this submission. Specifically, Academic AI (maintained by ACOmarket, the IT service provider for Austrian universities, and operated on an isolated Microsoft Azure instance; all inputs and uploaded documents are processed within the EU; user inputs are not used for model training; model used: OpenAI GPT-5; time of use: Q1/Q2 2026) was used to assist with language refinement. The authors critically reviewed, verified, and edited all AI-assisted content and accept full responsibility for the originality, accuracy, and scientific validity of the manuscript.

VII. REFERENCES

[1] Forman, J. L.; Poplin, G. S. et al. (2019): Automobile injury trends in the contemporary fleet: Belted occupants in frontal collisions. Traffic Injury Prevention, **20**(6): 607–612.

[2] Forman, J.; Jin, R.; Östling, M.; Jeppsson, H. (2025): Collision, Occupant, and Injury Characteristics in Low-to-moderate Speed Frontal Collisions – an Examination of U.S. Field Data from 2017-2022. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 10.-12.09.2025. Vilnius, Lithuania.

[3] Östling, M.; Jeppsson, H. et al. (2025): Injury Frequencies and Injury Risks in Frontal Car Crashes for AIS2+ and AIS1 Injuries that may lead to Long-Term Impairment – a GIDAS investigation. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 10.-12.09.2025. Vilnius, Lithuania.

[4] Forman, J. L.; Kent, R. W.; Mroz, K.; Pipkorn, B.; Bostrom, O.; Segui-Gomez, M. (2012): Predicting Rib Fracture Risk with Whole-Body Finite Element Models: Development and Preliminary Evaluation of a Probabilistic Analytical Framework. Annals of Advances in Automotive Medicine, **56**: 109–124.

[5] Larsson, K.-J.; Blennow, A.; Iraeus, J.; Pipkorn, B.; Lubbe, N. (2021): Rib Cortical Bone Fracture Risk as a Function of Age and Rib Strain: Updated Injury Prediction Using Finite Element Human Body Models. Frontiers in Bioengineering and Biotechnology, **9**.

[6] Forman, J. L.; Kulkarni, S.; Perez-Rapela, D.; Mukherjee, S.; Panzer, M.; Hallman, J. (2022): A Method for Thoracic Injury Risk Function Development for Human Body Models. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 14.-16.09.2022. Porto, Portugal.

[7] Iraeus, J.; Pipkorn, B. (2019): Development and Validation of a Generic Finite Element Ribcage to be used for Strain-based Fracture Prediction. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 11.-13.9. Florence, Italy.

[8] Iraeus, J.; Lindquist, M. (2020): Analysis of minimum pulse shape information needed for accurate chest injury prediction in real life frontal crashes. International Journal of Crashworthiness, **1**(2): 1–8.

[9] Euro NCAP. (2025) *Crash Protection Frontal Impact Protocol v1.1*.

[10] Euro NCAP. (2025) *Technical Bulletin Crash Protection 551: Human Body Model Simulation for Frontal Impacts v1.1*.

[11] Guleyupoglu, B.; Barnard, R.; Gayzik, F. S. (2017): Failed Rib Regions in a Computational Human Model vs. Observed PMHS Fractures: A Comparative Study. Ohio State University Injury Biomechanics Symposium,

[12] Toyota Motor Corporation (TMC); Toyota Central R&D Labs., Inc. (2021) *Documentation: Total Human Model for Safety (THUMS)*.

[13] Guleyupoglu, B.; Barnard, R.; Gayzik, F. S. (2017): Automating Regional Rib Fracture Evaluation in the GHBMC

Detailed Average Seated Male Occupant Model.

[14] Forman, J.; Gepner, B. et al. (2026): Strategies for Injury Prediction with Human Body Models. International Technical Conference on the Enhanced Safety of Vehicles. 12.-.15.05.2026. Toronto, Canada.

[15] Zhai, X.; Guo, Z. et al. (2019): High-speed X-ray visualization of dynamic crack initiation and propagation in bone. Acta Biomaterialia, **90**: 278–286.

[16] Euro NCAP. (2025) *Technical Bulletin Crash Protection 550: Qualification Procedure for Human Body Models v1.1*.

[17] Ressi, F.; Klug, C. (2024): Stochastic Replication of Injury Risks in Frontal Collisions with Different Human Body Models in a Generic Vehicle Interior. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 11.-13.9. Stockholm, Sweden.

[18] Open Virtual Testing Organisation (OVTO) (2025): HBM4VT Validation Catalogue, https://openvt.eu/EuroNCAP/hbm4vt-validation-catalogue.

[19] Ressi, F.; Schneider, B. et al. (2025): Development of a Robust Human Body Model Qualification Methodology: Ensuring Biofidelity for Occupant Protection Assessments in Virtual Testing. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 10.-12.09.2025. Vilnius, Lithuania.

[20] Bouquet, R.; Ramet, M.; Bermond, F.; Cesari, D. (1994): Thoracic and pelvis human response to impact. International Technical Conference on the Enhanced Safety of Vehicles. 23.-25.05.1994. Munich, Germany.

[21] 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.

[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] 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.

[24] 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.

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

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

[27] Viano, D. C. (1989): Biomechanical Responses and Injuries in Blunt Lateral Impact. Stapp Car Crash Conference. 4.-6.10.1989. Washington, D.C., USA.

[28] 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.

[29] Rubin, D. B. (1981): The Bayesian Bootstrap. The Annals of Statistics, **9**(1).

[30] Open Virtual Testing Organisation (OVTO) (2026): Dynasaur+, https://openvt.eu/dynasaur/dynasaur-plus/-/tree/8-order-of-calculations-for-strains.

[31] Fahlstedt, M.; Meng, S.; Kleiven, S. (2022): Influence of Strain post-processing on Brain Injury Prediction. Journal of biomechanics, **132**: 110940.

[32] Blade, S.; von Kleeck III, W.; Koya, B.; Mischo, S.; Devane, K.; Gayzik, S. F. (2024): Meta-Analysis of Rib Fracture in GHBMC Detailed Human Body Models: A Retrospective Study. IRCOBI Conference. International Research Council on the Biomechanics of Injury. 11.-13.9. Stockholm, Sweden.

[33] Pipkorn, B.; Niranjan Poojary, Y.; Östh, J.; Larsson, K.-J.; Iraeus, J. (2026): Assessing Injury Prediction with Human Body Models: Application to Rib Fracture Risk Prediction Using SAFER HBM. SAE International Journal

of Transportation Safety, **14**(1).

[34] Duma, S.; Stitzel, J.; Kemper, A.; McNally, C.; Kennedy, E.; Matsuoka, F. (2005): Non-Censored Rib Fracture Data from Dynamic Belt Loading Tests on The Human Cadaver Thorax. International Technical Conference on the Enhanced Safety of Vehicles. 6.-9.6.2005. Washington, D.C., USA.

VIII. APPENDIX

Boxplots of maximum strains for all 24 ribs for each HBM

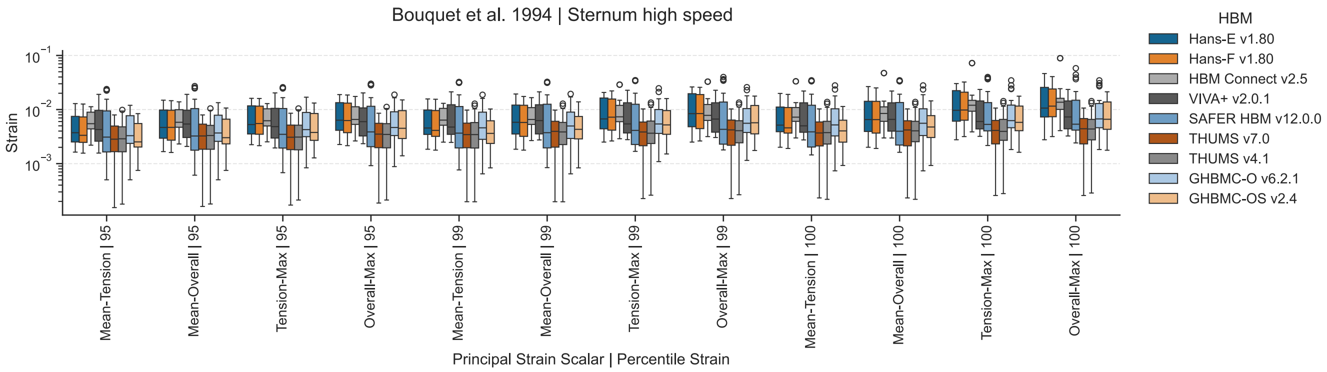

Fig. A1. Boxplots of maximum strains for all 24 ribs for each HBM in load case Bouquet *et al*. 1994, Sternum high speed. Note that a logarithmic axis was used for the strains.

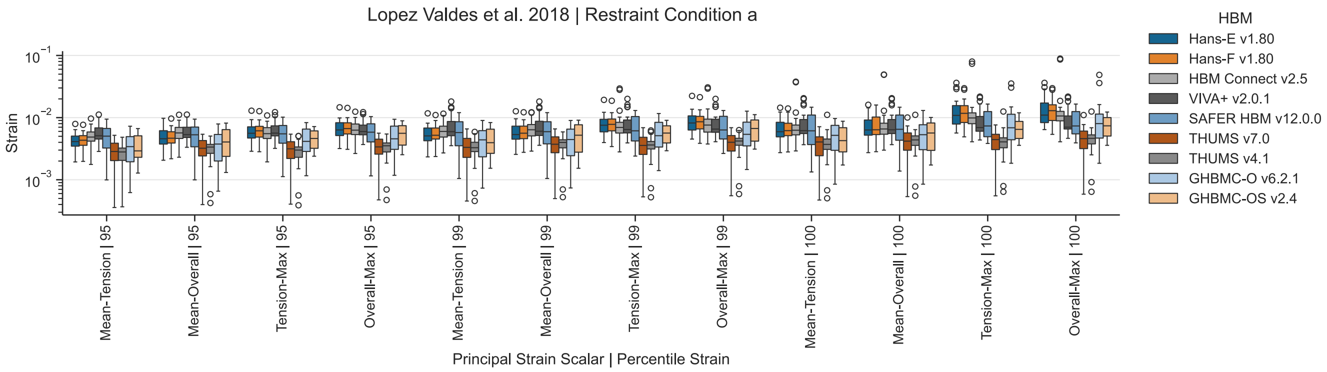

Fig. A2. Boxplots of maximum strains for all 24 ribs for each HBM in load case Lopez Valdes *et al*. 2018, Restraint Condition a. Note that a logarithmic axis was used for the strains.

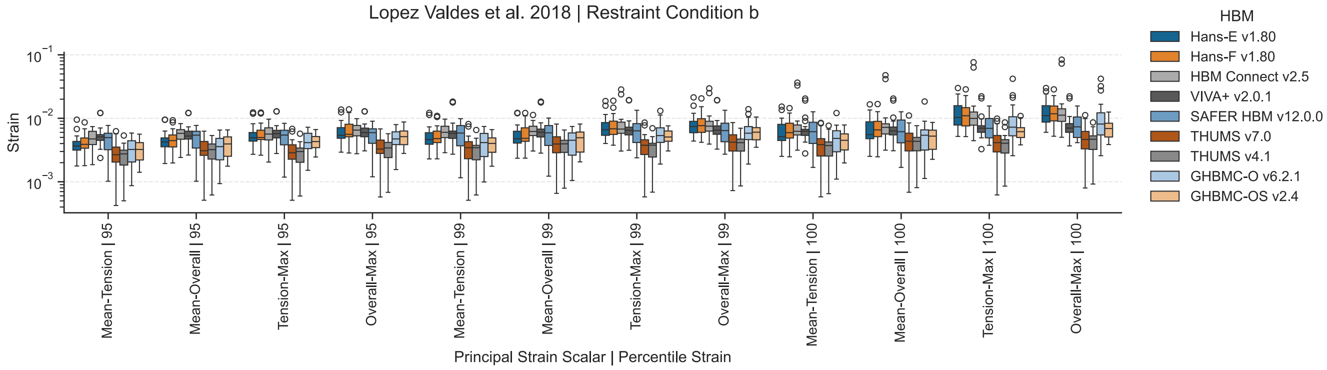

Fig. A3. Boxplots of maximum strains for all 24 ribs for each HBM in load case Lopez Valdes *et al*. 2018, Restraint Condition b. Note that a logarithmic axis was used for the strains.

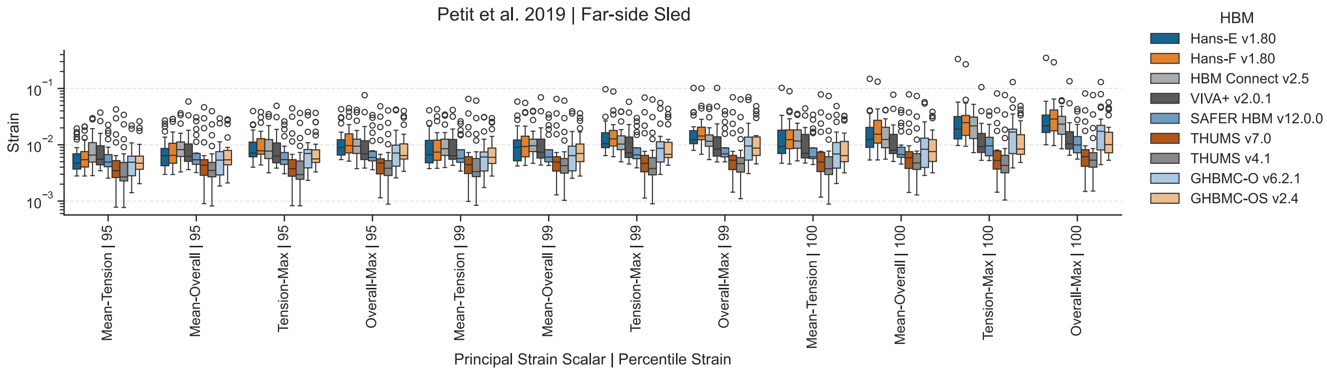

Fig. A4. Boxplots of maximum strains for all 24 ribs for each HBM in load case Petit *et al*. 2019, Far-side Sled. Note that a logarithmic axis was used for the strains.

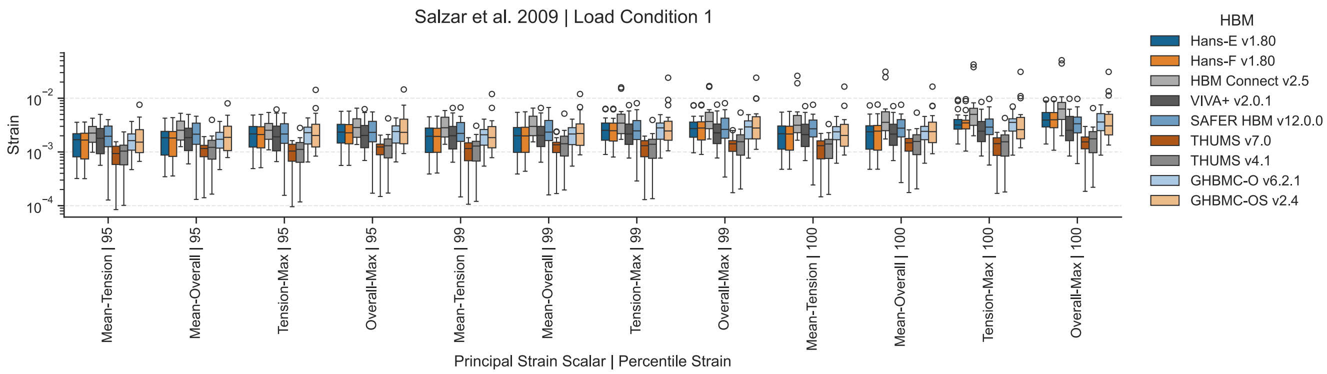

Fig. A5. Boxplots of maximum strains for all 24 ribs for each HBM in load case Salzar *et al.* 2009, Load Condition 1. Note that a logarithmic axis was used for the strains.

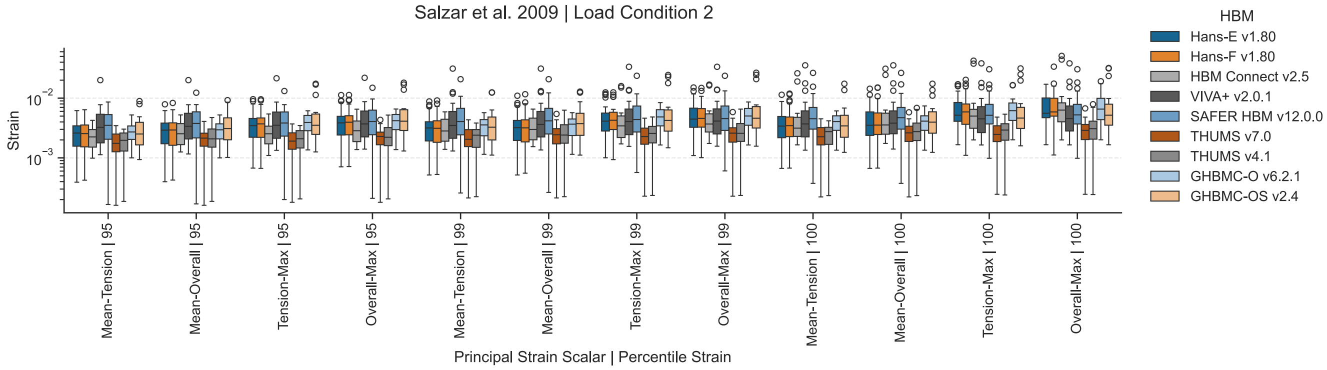

Fig. A6. Boxplots of maximum strains for all 24 ribs for each HBM in load case Salzar *et al*. 2009, Load Condition 2. Note that a logarithmic axis was used for the strains.

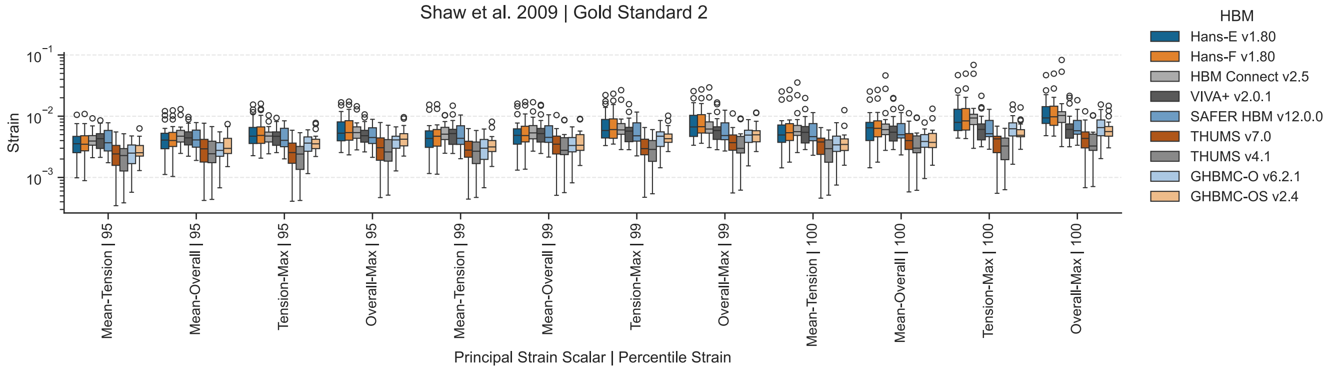

Fig. A7. Boxplots of maximum strains for all 24 ribs for each HBM in load case Shaw *et al*. 2009, Gold Standard 2. Note that a logarithmic axis was used for the strains.

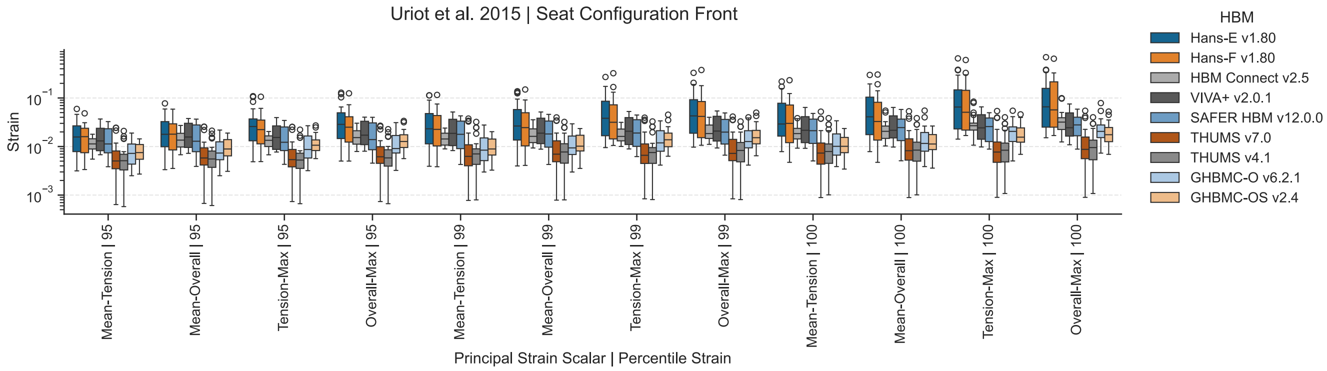

Fig. A8. Boxplots of maximum strains for all 24 ribs for each HBM in load case Uriot *et al*. 2015, Seat Configuration Front. Note that a logarithmic axis was used for the strains.

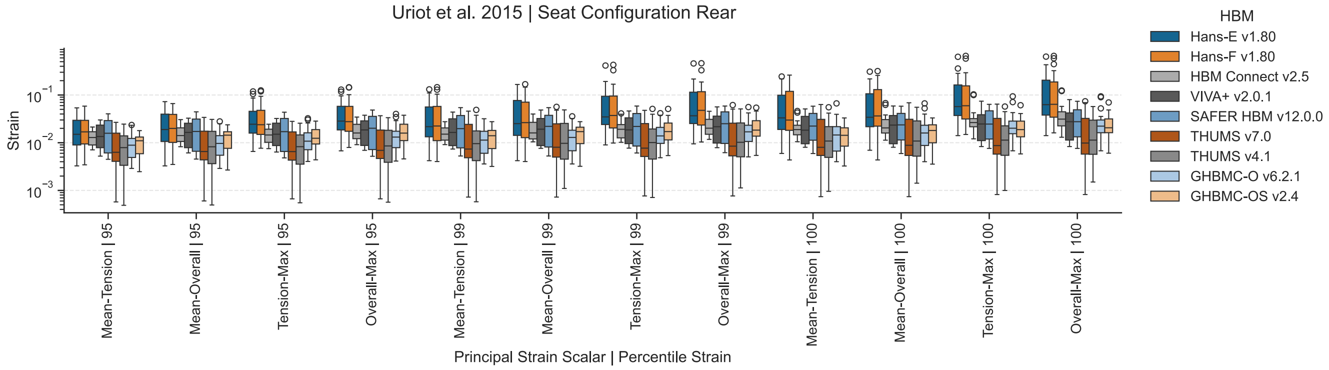

Fig. A9. Boxplots of maximum strains for all 24 ribs for each HBM in load case Uriot *et al*. 2015, Seat Configuration Rear. Note that a logarithmic axis was used for the strains.

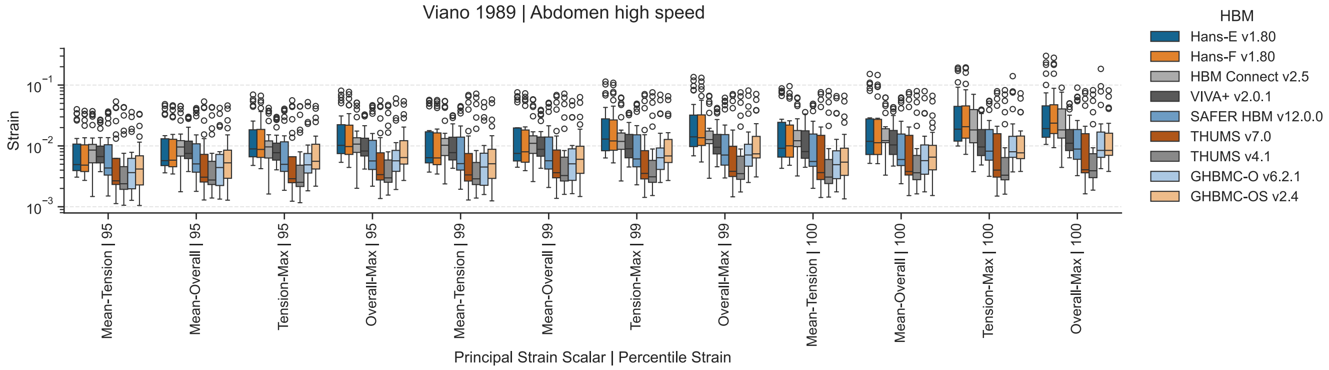

Fig. A10. Boxplots of maximum strains for all 24 ribs for each HBM in load case Viano 1989, Abdomen high speed. Note that a logarithmic axis was used for the strains.

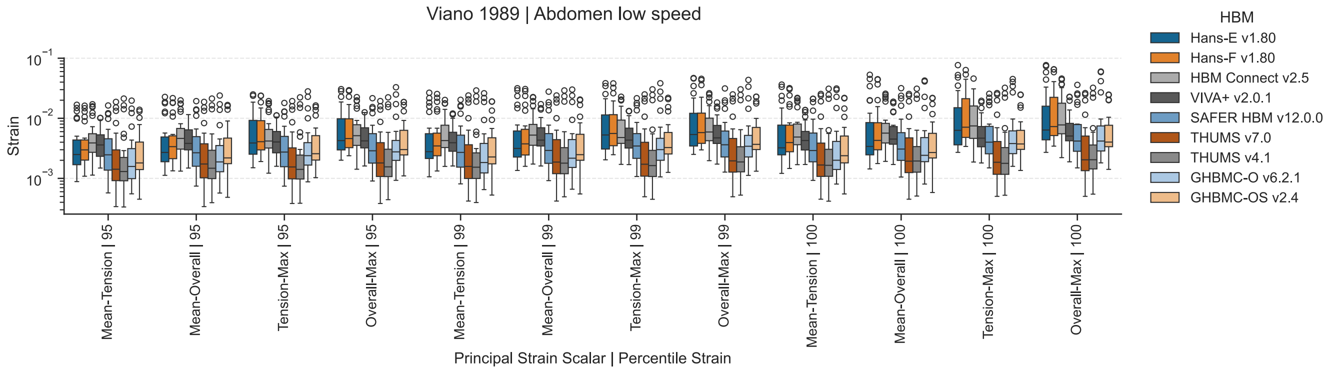

Fig. A11. Boxplots of maximum strains for all 24 ribs for each HBM in load case Viano 1989, Abdomen low speed. Note that a logarithmic axis was used for the strains.

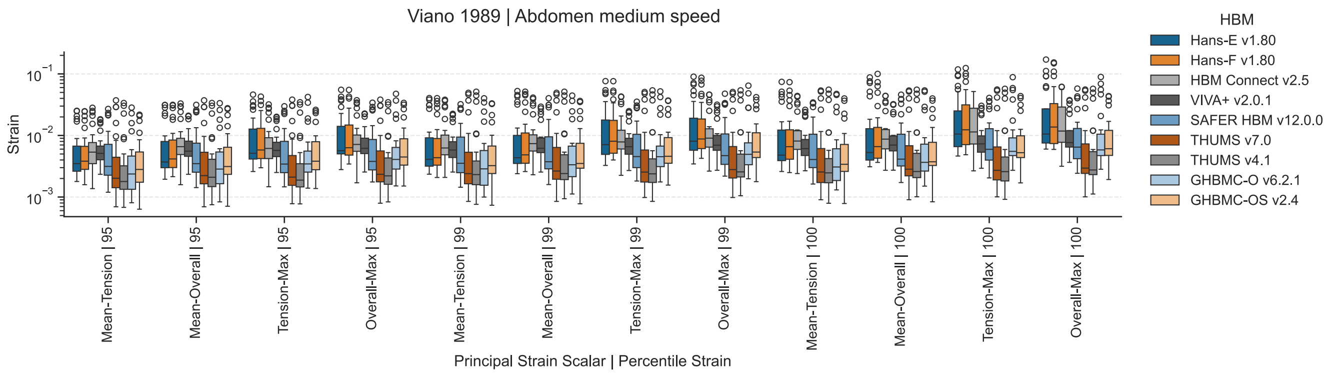

Fig. A12. Boxplots of maximum strains for all 24 ribs for each HBM in load case Viano 1989, Abdomen medium speed. Note that a logarithmic axis was used for the strains.

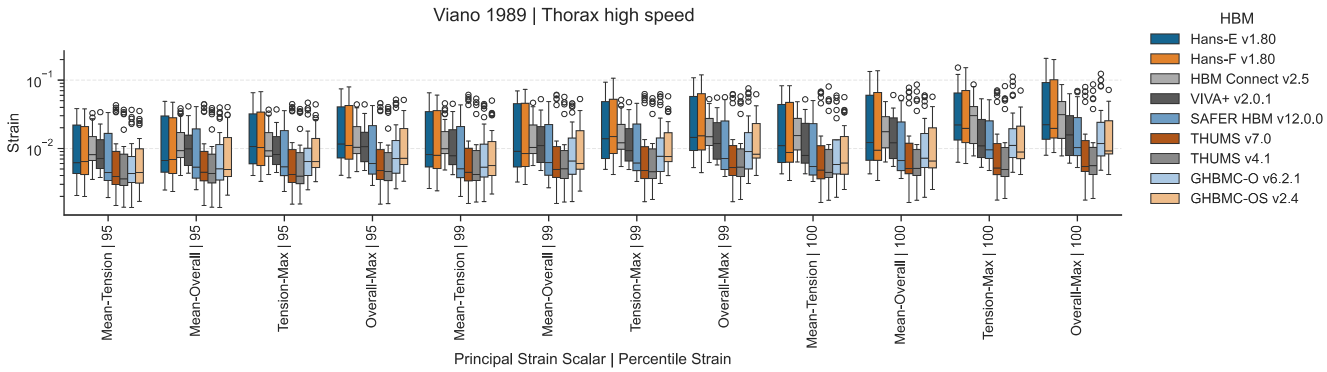

Fig. A13. Boxplots of maximum strains for all 24 ribs for each HBM in load case Viano 1989, Thorax high speed. Note that a logarithmic axis was used for the strains.

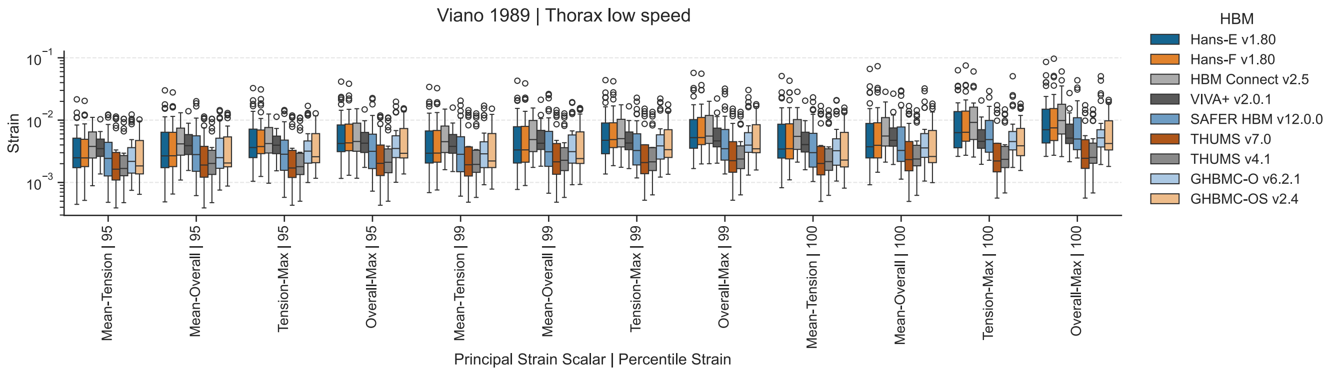

Fig. A14. Boxplots of maximum strains for all 24 ribs for each HBM in load case Viano 1989, Thorax low speed. Note that a logarithmic axis was used for the strains.

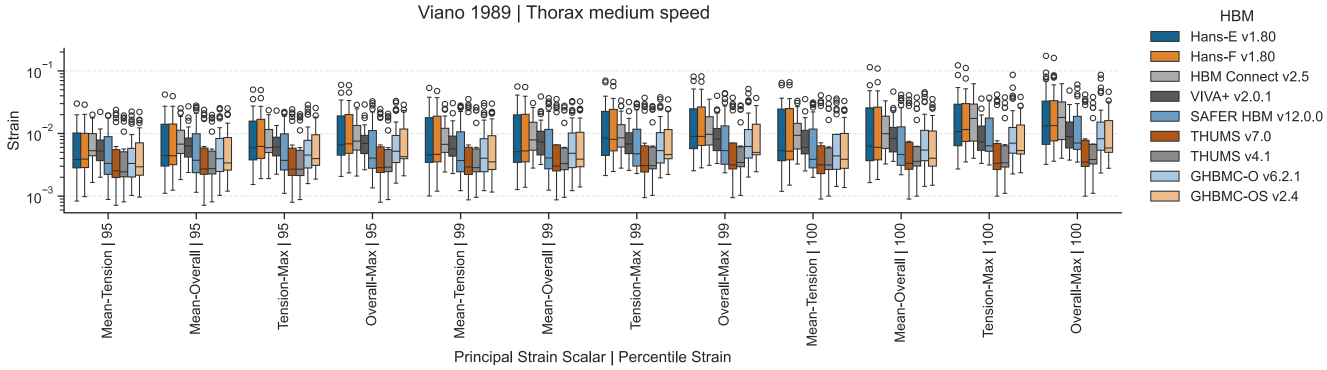

Fig. A15. Boxplots of maximum strains for all 24 ribs for each HBM in load case Viano 1989, Thorax medium speed. Note that a logarithmic axis was used for the strains.

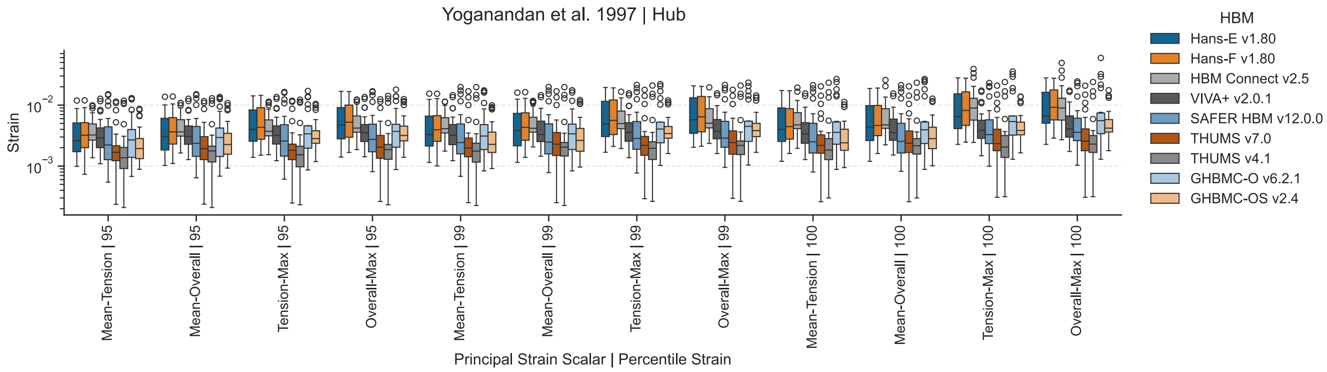

Fig. A16. Boxplots of maximum strains for all 24 ribs for each HBM in load case Yoganandan *et al.* 1997, Hub. Note that a logarithmic axis was used for the strains.

NFR3+ risk calibration diagnostics by load case comparing PMHS outcomes and HBM predictions

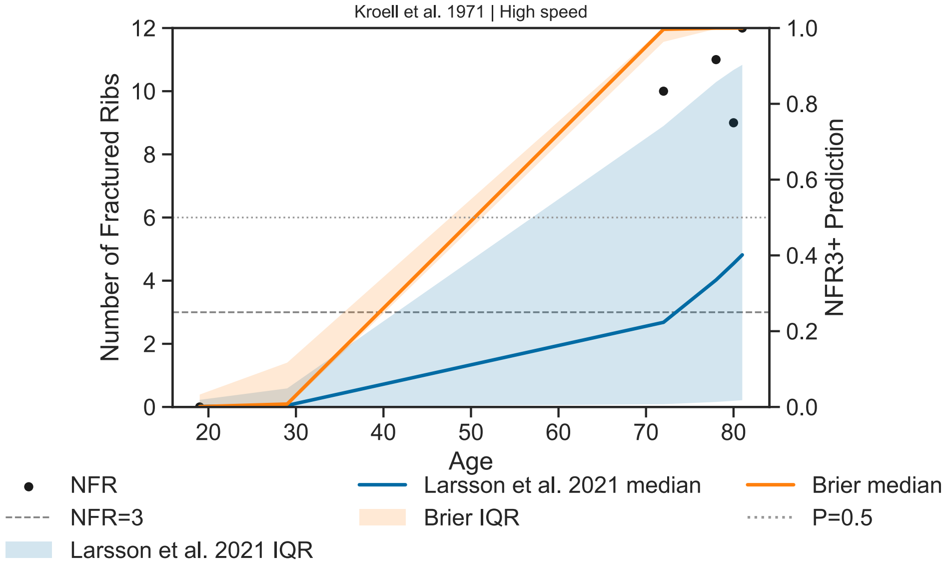

Fig. A17. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Kroell *et al.* 1971. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

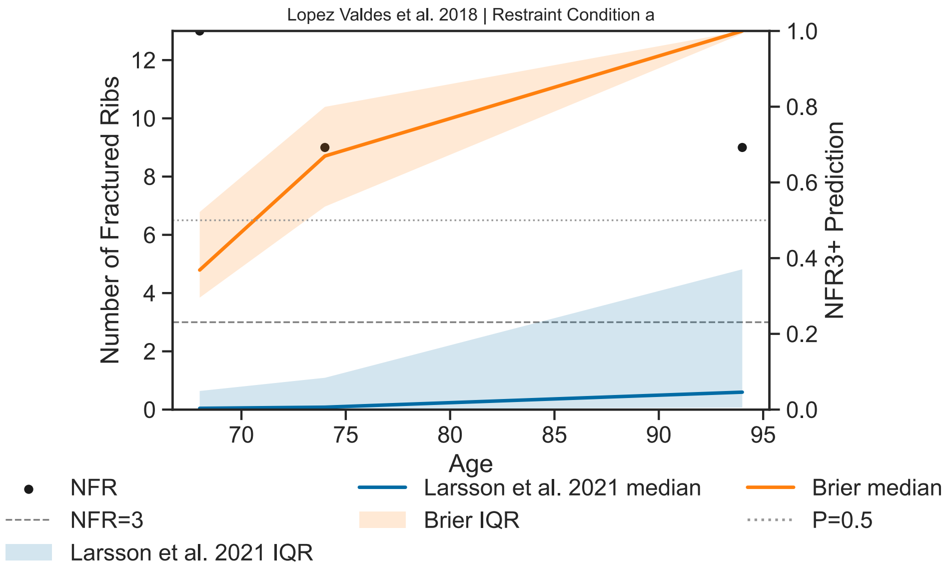

Fig. A18. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Lopez Valdes *et al.* 2018 RCa. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

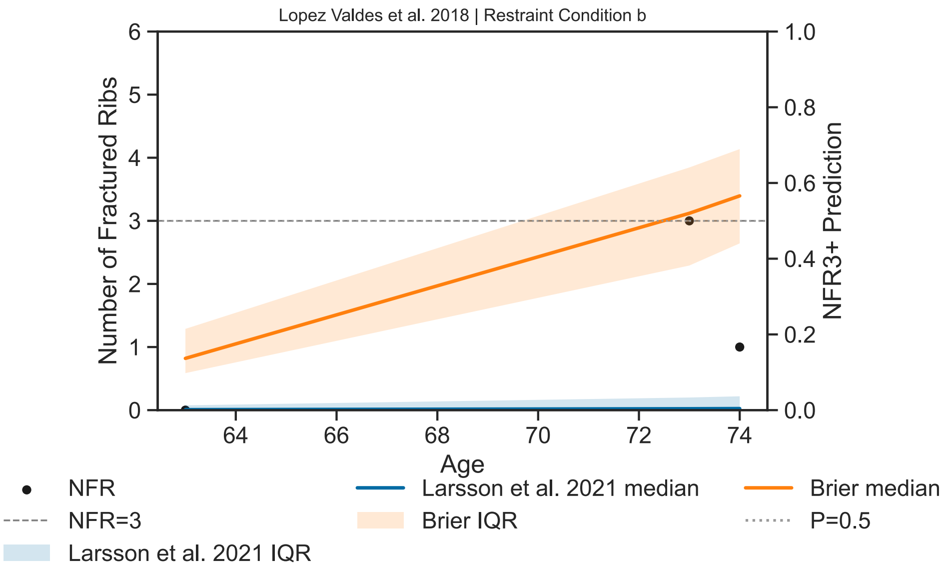

Fig. A19. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Lopez Valdes *et al.* 2018 RCb. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

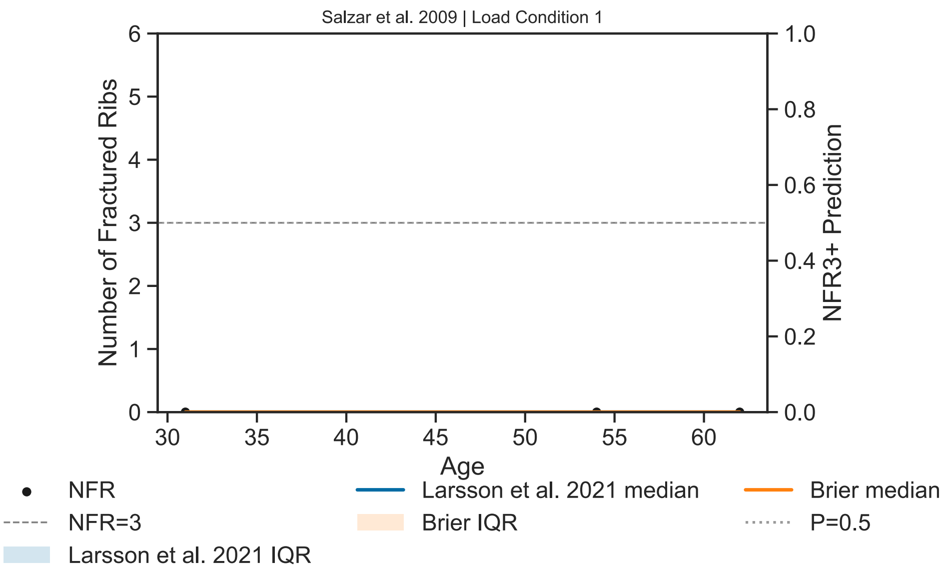

Fig. A20. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Salzar *et al.* 2009 LC1. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

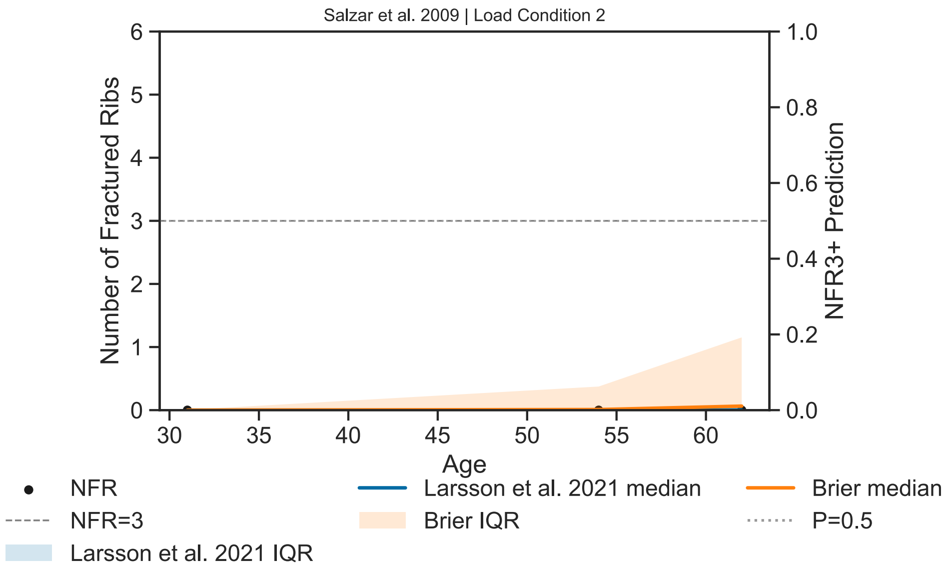

Fig. A21. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Salzar *et al.* 2009 LC2. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

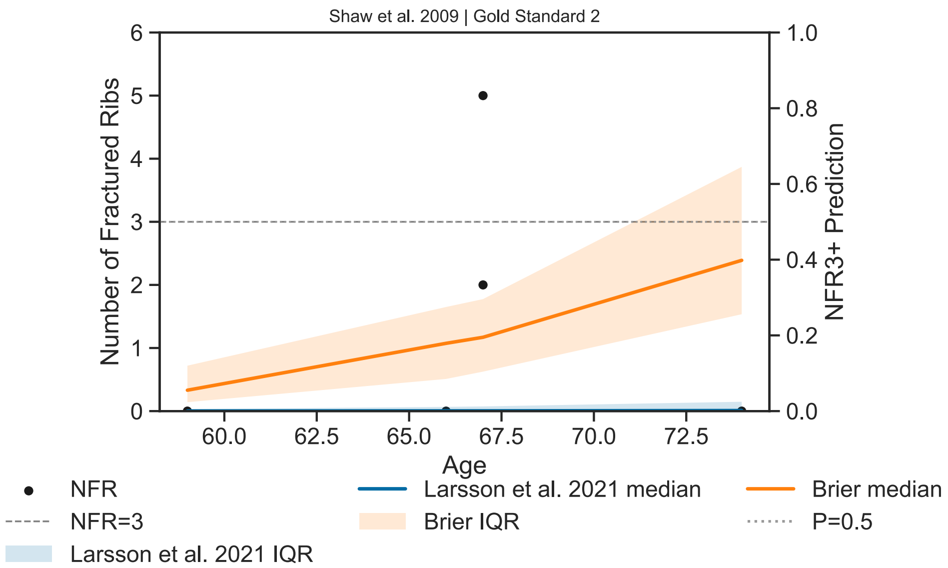

Fig. A22. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Shaw *et al.* 2009 Gold Standard 2. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

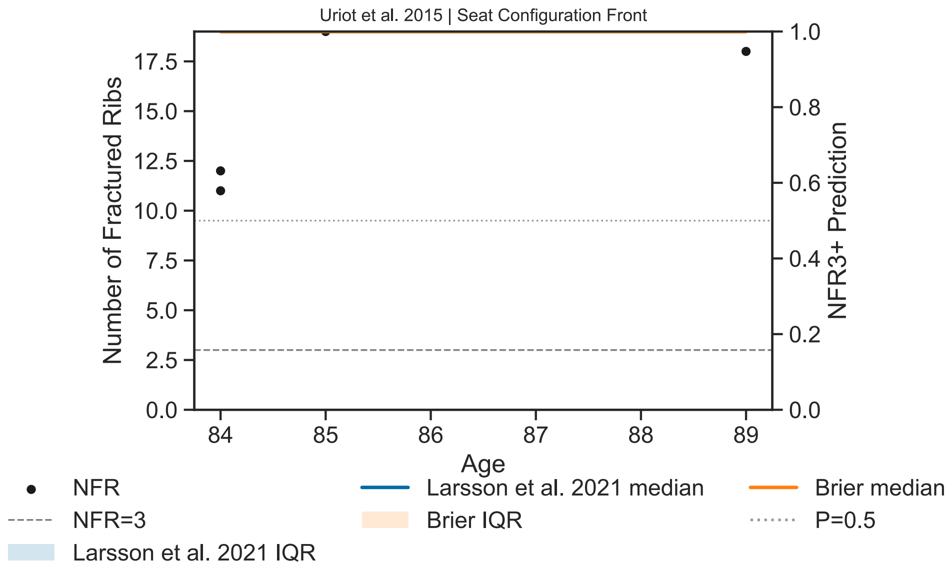

Fig. A23. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Uriot *et al.* 2015 SCfr. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

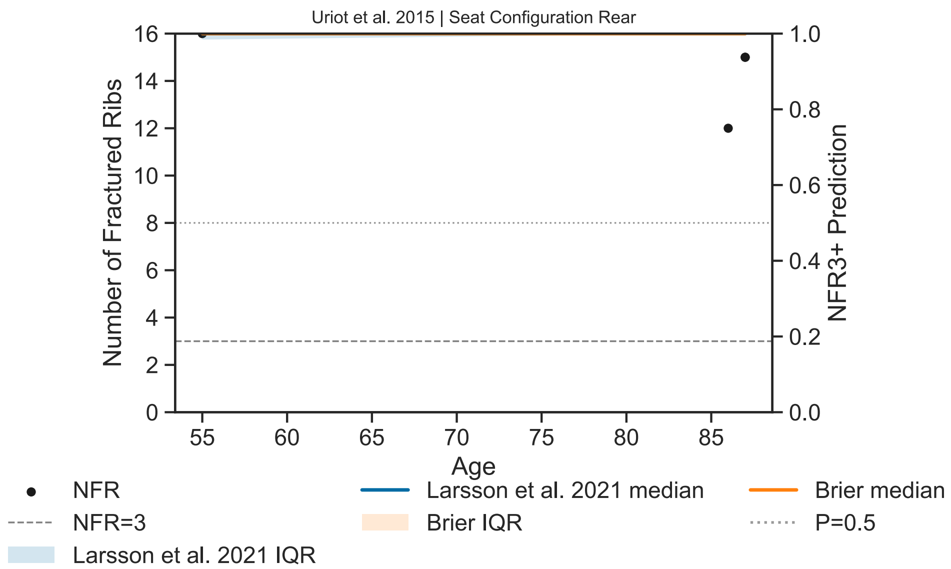

Fig. A24. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Uriot *et al.* 2015 SCre. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

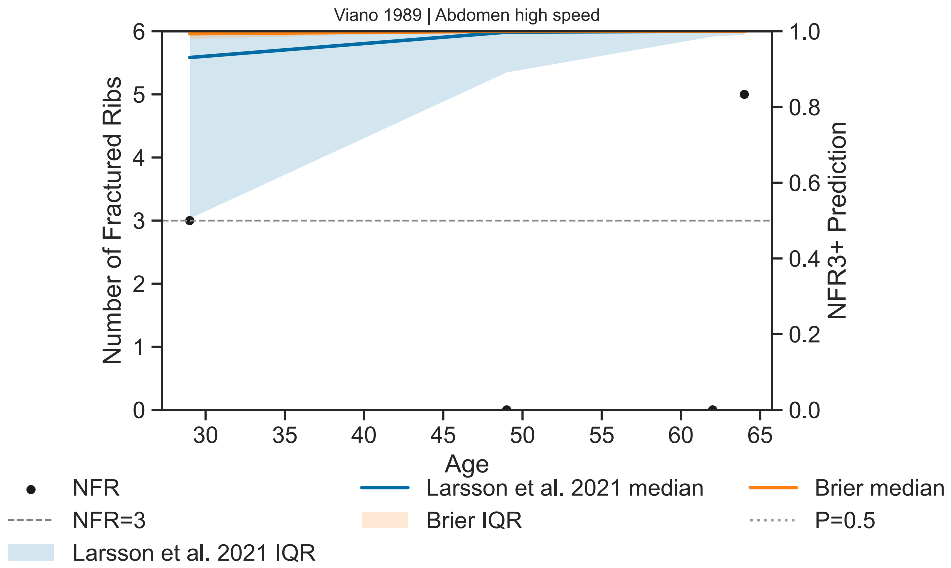

Fig. A25. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Viano 1989 Abdomen high speed. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

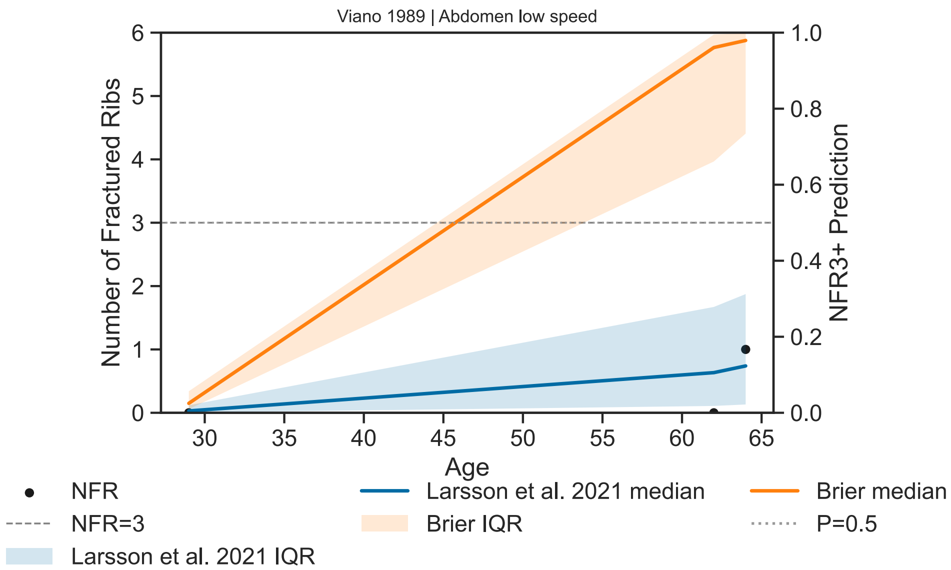

Fig. A26. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Viano 1989 Abdomen low speed. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

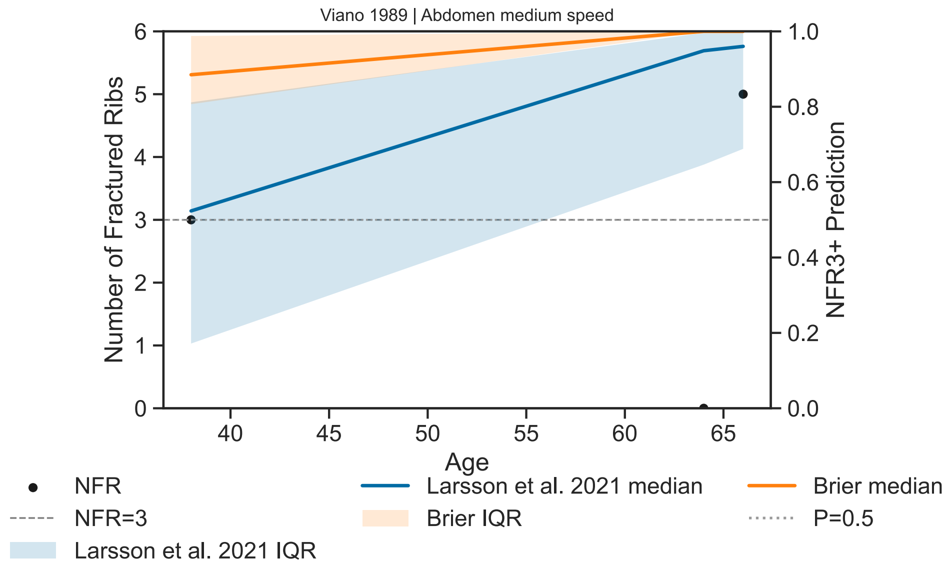

Fig. A27. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Viano 1989 Abdomen medium speed. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

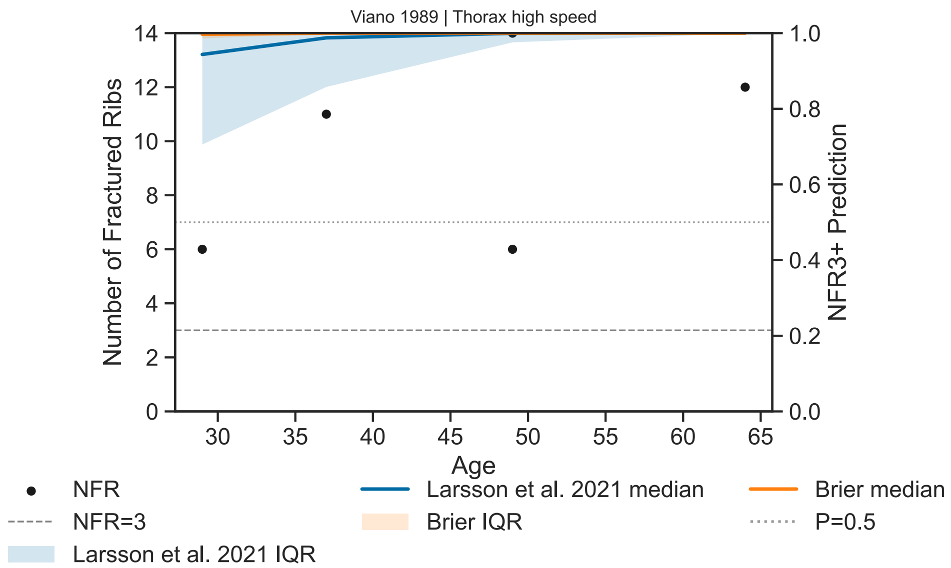

Fig. A28. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Viano 1989 Thorax high speed. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

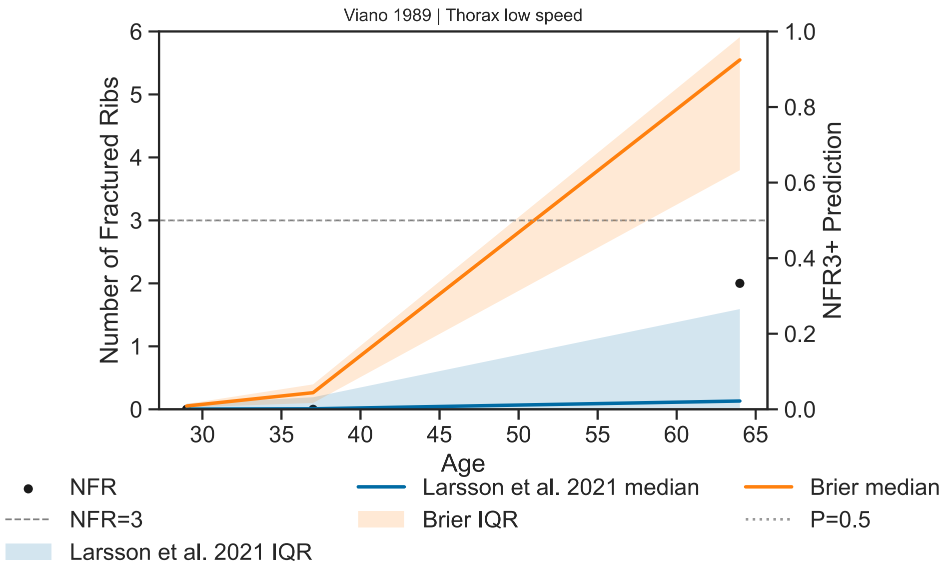

Fig. A29. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Viano 1989 Thorax low speed. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

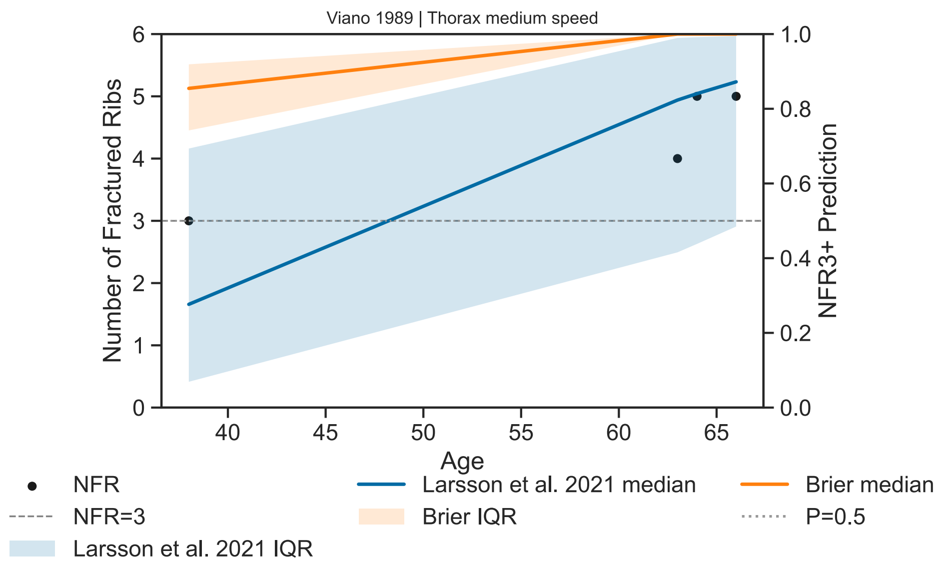

Fig. A30. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Viano 1989 Thorax medium speed. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

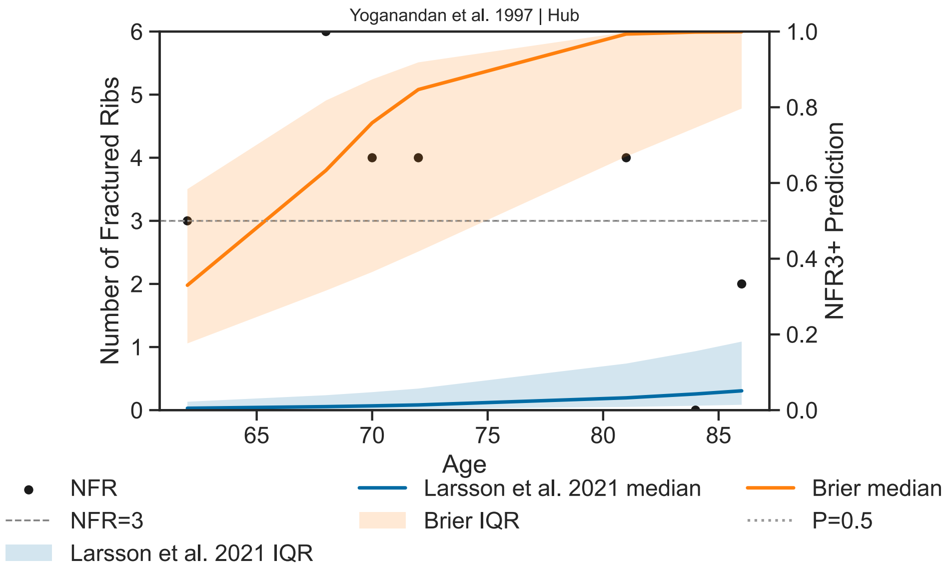

Fig. A31. Plots showing the number of fractured ribs in PMHS tests as function of subject age (black dots) for Yoganandan *et al.* 1997. Resulting mean and interquartile range of HBM NFR3+ risk predictions is shown for the original IRF (blue) and calibrated IRF (based on Brier approach). The dashed horizontal line indicates three rib fractures as the relevant threshold for NFR3+.

Bar charts illustrating calibration results for different strategies for all HBMs

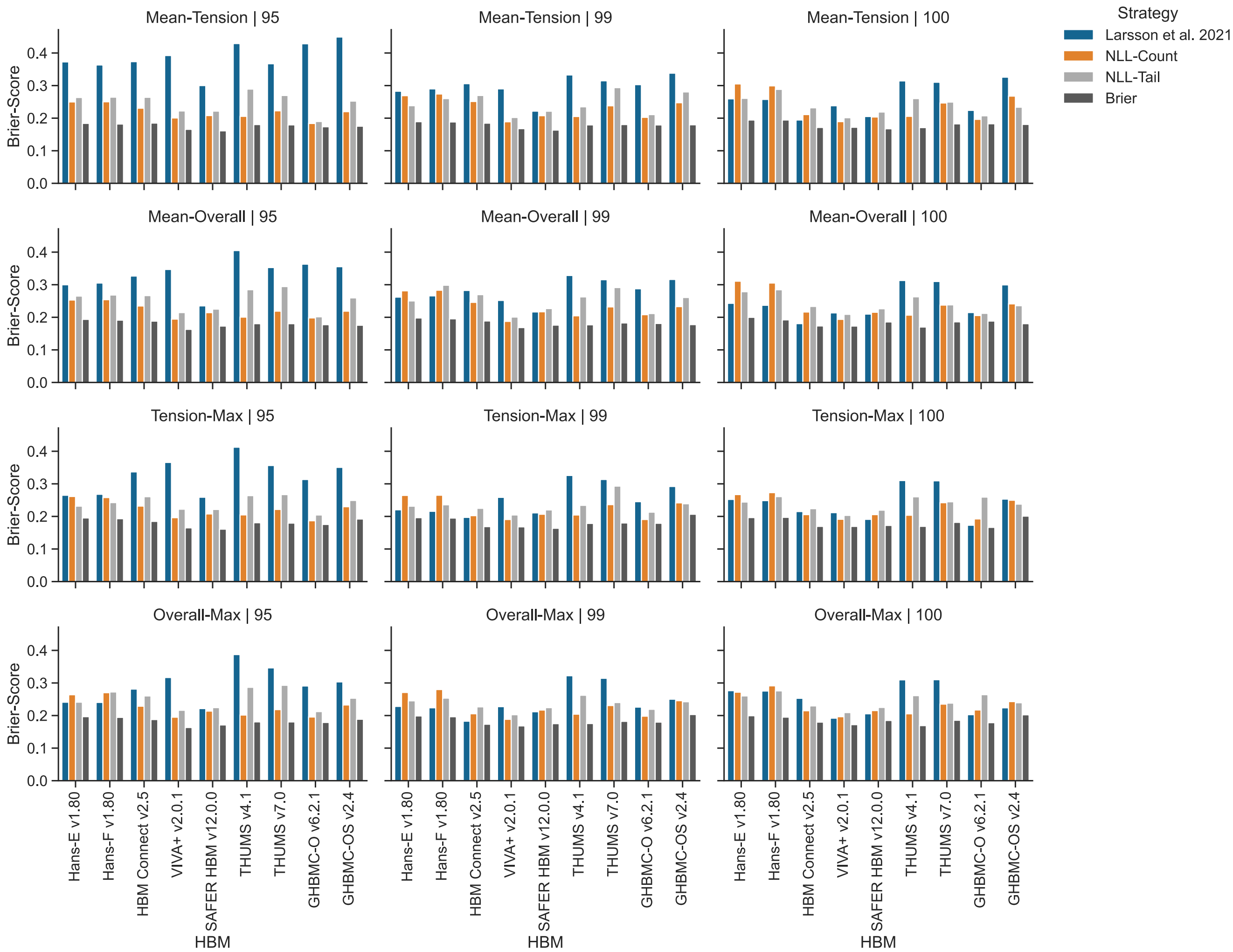

Fig. A32. Comparison of calibration results for the three strategies and the baseline (Larsson *at al.* 2021 [5]) three. Brier scores (lower is better) are provided, regardless if Brier score was used as calibration target. The 12 plots show all combinations of strain percentiles and principal strain scalar options.

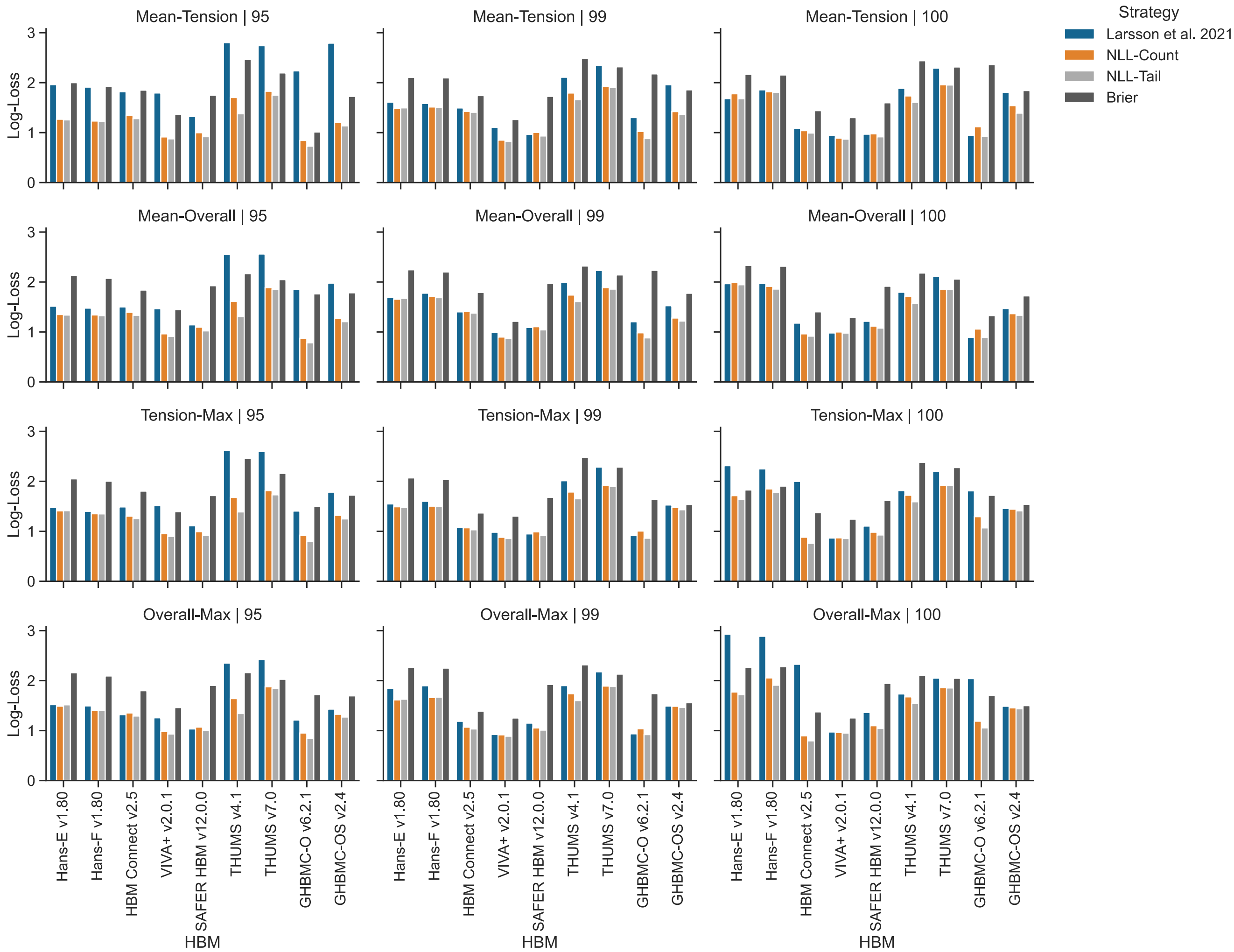

Fig. A33. Comparison of calibration results for the three strategies and the baseline (Larsson *at al.* 2021 [5]) three. Log-loss values (lower is better) are provided, regardless whether log-loss was used as calibration target. The 12 plots show all combinations of strain percentiles and principal strain scalar options.

Boxplots illustrating calibration results for different strategies aggregated over all HBMs

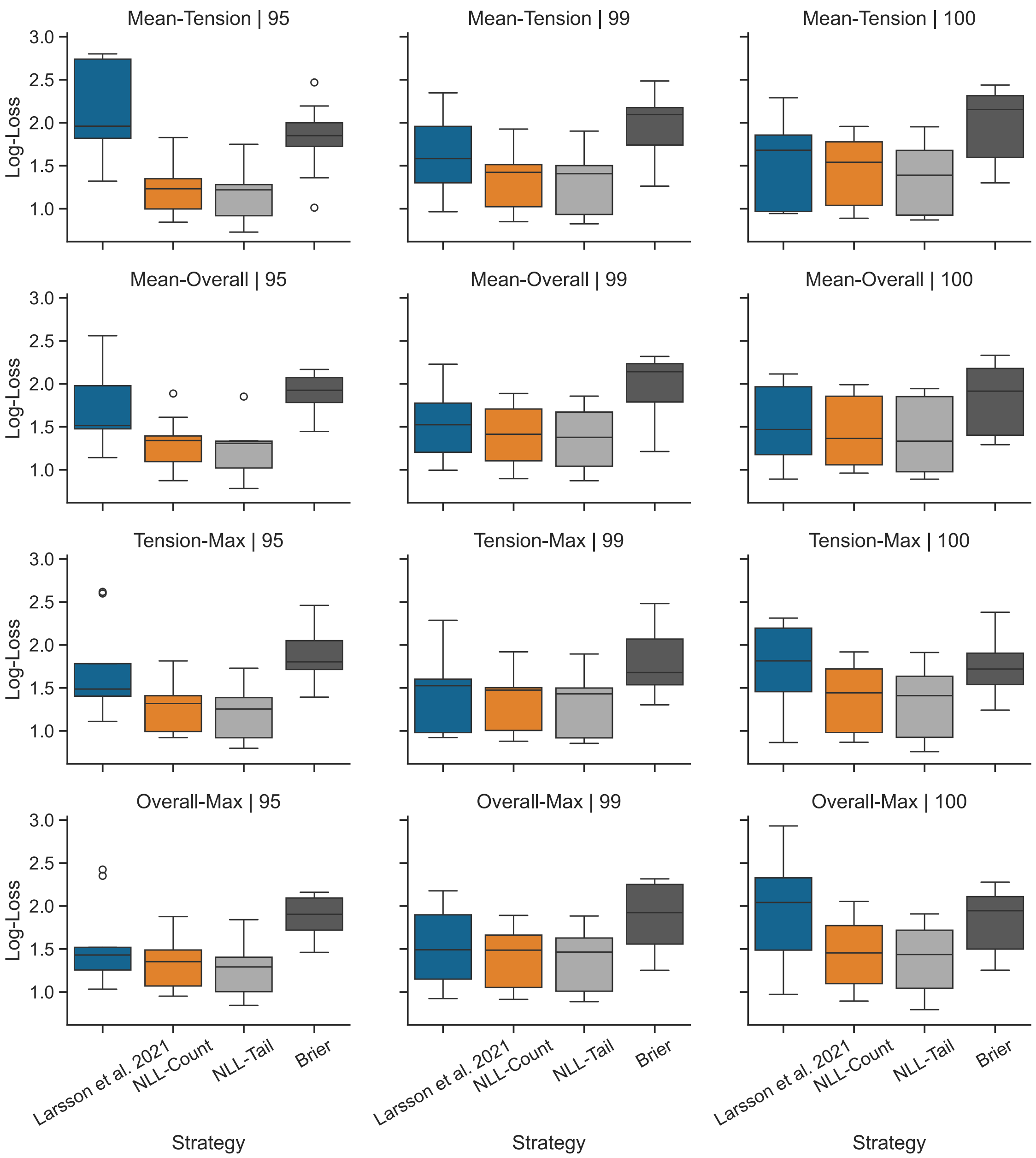

Fig. A34. Comparison of calibration results for the three strategies and the baseline (Larsson *at al.* 2021 [5]) with the results of the nine HBMs aggregated in box plots. log-loss values (lower is better) are provided, regardless whether log-loss was used as calibration target. The 12 box plots show all combinations of strain percentiles and principal strain scalar options.

Weibull CDFs of single rib fracture risk as a function of strain for all HBMs

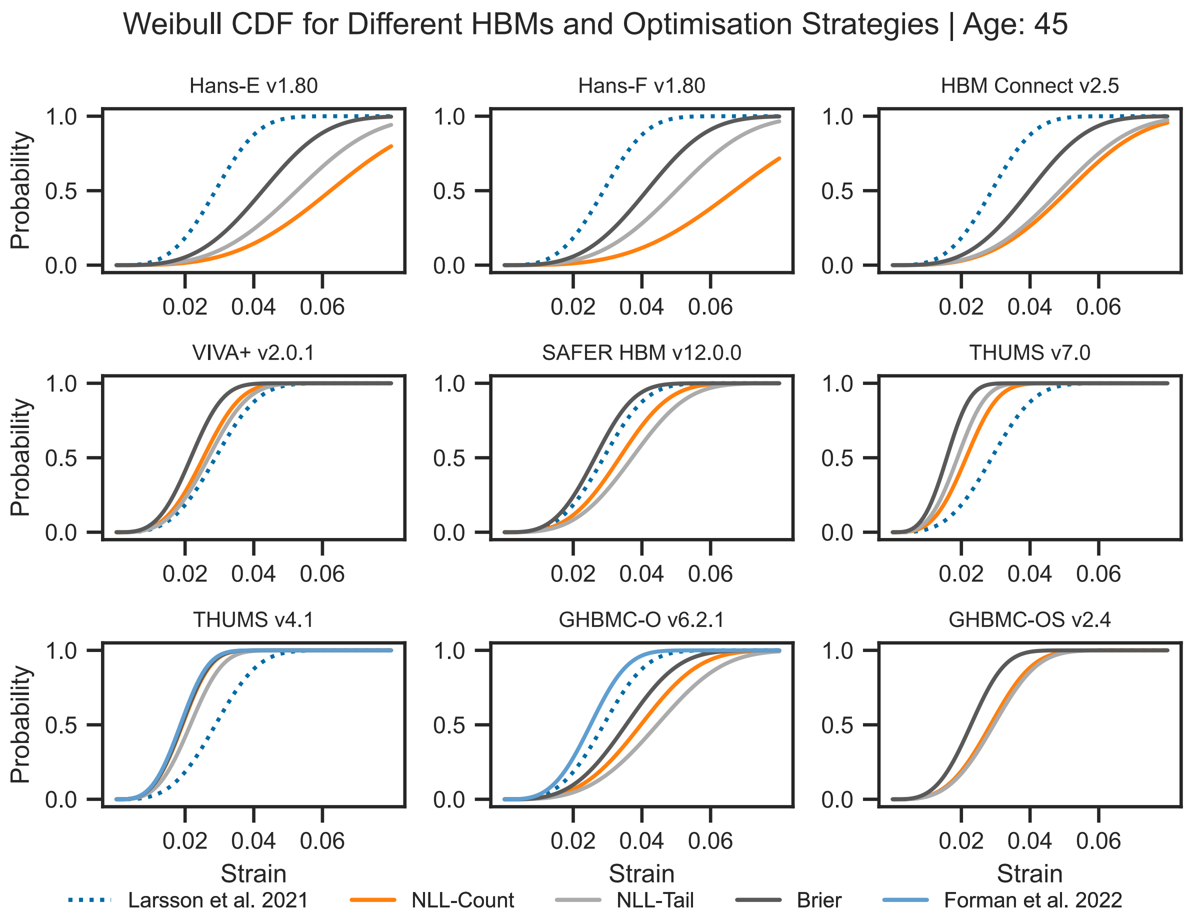

Fig. A35. Weibull CDFs of single rib fracture risk as a function of strain for all HBMs for an age of 45. Each subplot overlays the IRFs fitted via different optimisation strategies (NLL-Count, NLL-Tail, and Brier) with the baseline IRF from [5]. Model-specific Weibull CDFs are added from [6] where applicable.