Peer Review Response to Comments
Injury Criteria for the THOR 50th Male ATD

The National Highway Traffic Safety Administration plans to release the technical report, "Injury Criteria for the THOR 50th Male ATD". This report documents the process by which injury criteria were developed and/or selected for the THOR-50M and provides other details related to these efforts.

Overall Reviewer Recommendations:

Reviewer	Recommendation
Reviewer 1 (R1)	Acceptable with minor revision
R2	Acceptable with minor revision
R3	"several major concerns"
R4	Acceptable with minor revision
R5	Acceptable with major modification
R6	Acceptable with recommended and required revisions, ranging from minor to major

Response to Specific Charge Questions:

1. Are the analytical methods and data used to create the proposed injury criteria and limits appropriate?

    R1: Generally speaking, nearly all of the described methods selected for creating the injury criteria are subject to limitations and could be contested by competing approaches. However, the selected injury criteria are well supported by historical, peer-reviewed research that has been accepted by the scientific injury biomechanics community.

    R1: In nearly all cases, limitations do exist with the criteria. Those limitations are either specifically stated or alluded to with further explanations of additional data and examples. In cases where supporting research was unavailable to address a limitation, historical precedent and norms were employed.

    R2: In general, the analytical methods and data used to create the proposed injury criteria and limits are appropriate. In my opinion, the report includes a very thorough review and comparison to all available data sources, M&S comparisons, test results and the data is appropriately compared to the derived injury criteria for the THOR.

    R4: The analytic methods and data used to create the proposed injury criteria and the limits that are provided in draft document are generally appropriate.

2. Is the organization of the document appropriate and does it present the material in a clear and concise manner?

    R1: The report is generally well structured, clearly written and nicely balanced in citing both supporting literature and known limitations.

R1: The organization by anatomy and then subsequently by field data, literature review, design, etc. is logical and produces the appropriate progression of information.
R2: The layout of the material is clearly organized and appropriate for leveraging the existing data. I appreciated the grouping by body region and the fact that each section covered the available field/ data, literature review, ATD design, biofidelity, injury formulation and fleet data with the THOR-50M ATD. I thought that was a logical progression.
R4: The organization of the document is appropriate and the presentation is clear. I have some minor suggestions on how presentation and organization could be improved that are detailed in my recommendations, below. Presenting information in a concise manner in a document such as this is difficult and maybe counter-productive as NHTSA's injury criteria documents are meant to have long-term reference value and thus a complete presentation (including data) is desirable.

3. In your opinion, what are the weakest and strongest parts of the technical report? Please make suggestions on how the weakest parts of the report can be addressed.

   R1: A number of my comments above imply 'strong' and 'weak' aspects of this report. A few higher level comments are repeated here:
   - Strongest:
     - The layout and content of the document, ranging from field data to detailed biomechanical studies, permits this to be somewhat of a standalone document that provides a summary of the necessary detail to understand the origins of a given criteria.
     - The selected injury criteria are rooted in decades of research but they are balanced by the identification of limitations.
   - Weakest:
     - Matched pair testing should be conducted, where sensible, for all injury criteria. If the goal of this document is to provide injury criteria as applied to THOR 50M measurements, then the matched pair tests should be addressed prior to finalizing this document.

   R2: Strongest: I felt the strongest part of the technical report was the statistical assessment methodology and the human injury risk formulation. The statistical methods are thorough and current. And the assessment of significant variables and the treatment of co-variates were well illustrated.

   R2: Weakest: In my opinion, the areas of the technical report that left some lingering questions were the methods to translate the PMHS criteria to the THOR-50M. Traditionally, or maybe it is better stated as "ideally" a matched-pair test series (where the ATD component is exposed to the same individual test conditions of every PMHS that was included in the creation of the injury risk curve) would allow for more direct translation and creation of a risk curve specific to that ATD. In the appendix, it looks as if some series were match paired tested, but others were not. Looking at Table S1 (Summary of THOR-50M Injury Measures and Values at 10, 25, and 50%

risk), and comparing that to Figure 8.8 (Logistic regression of upper tibia axial force data, including mean injury risk function and 95% confidence intervals, assuming mass of 75kg) it appears the recommend THOR-50M criteria for upper tibia force (N) is the same as the PMHS probability of injury. I understand the lower leg demonstrated good or excellent biofidelity in the three primary conditions: dynamic heel impact, dynamic axial compression and dynamic dorsiflexion. However, completing the exact match pair test conditions with the ATD leg, scaling or transforming the data with the corresponding specimen influential factors, and completing the survival analysis to generate a THOR-specific risk curve could provide even greater accuracy in the injury assessment during a vehicle evaluation. If I have mis-read the report and there was, in fact, matched pair testing completed to directly transform the PMHS data to THOR-50M application, I must have missed it. That may be an area that could use extra emphasize.

R2: Another area that may warrant additional details in this technical report regarding the injury criteria for the THOR-50M is the topic of ATD data transformation and filtering. If that topic has been covered in other technical reports, it may be worth including in the appendix or referencing the in body of the paper. Considering the fact that the ATD load cells may be in different location to the PMHS (requiring possible geometric and kinematic transformation) or considering simple human variation (which could drive transformations for length or mass compensation, etc.), those translations could noticeable shift the risk curve values, thus having significant impact on a vehicle assessment. Looking at Figure 9.2 and Figure 9.3 (Full Frontal and Oblique fleet predicted injury risk with a driver seated THOR-50M versus field estimates and associated 95% confidence intervals), it seems that the THOR-50M injury thresholds significantly over predicts injuries in all body regions except the lower extremities. I would look to the matched pair transformations as a prime opportunity to reduce the difference in injury prediction and actual occurrence.

Similarly, the data filtering is an inherent part of the injury assessment values. There are many diverse opinions on what frequency content is necessary and how much should be retained. As a minimum, it would be good to include what filter classes are recommended for ATD usage and mention, if matched pair testing was executed that the same filtering was performed for the generation of the THOR-50M's injury reference curves.

R3: In general, I found the report to be well organized and clear in its formulation of the injury criteria.

R3: I think there can be a little more specificity as noted in the comments below.

R4: This report has many strengths. The level of detail in the analyses used to develop BrIC and address concerns with the original formulation and the work to explore the effects of BrIC on restraint system optimization are excellent. The inclusion of the data used to develop injury risk curves and the clear explanations for the rationale for the choices made in risk curve developed are also major strengths.

R4: The weaknesses of the report are the limitations already acknowledged by NHTSA in each of the chapters. Almost all of these are associated with dummy biofidelity (e.g., ankle inversion/eversion) or lack of sufficient experimental data on which to base injury criteria (e.g., muscle effects in multiple body regions).

Specific Comments:

Summary

R1: The Summary should begin with a statement of the report's purpose/scope.

Response: An "Overview" section was added to the beginning of the summary.

R1: A single table summarizing the determined THOR 50th injury criteria would provide a nice complement to the Summary. It would be an artifact that could be used beyond its inclusion within this report.

Response: Good suggestion; a table was added to the end of the Summary.

R1: The anatomical sub-sections within the Summary are inconsistent in their inclusion of limitations. For example, the neck section includes 'As the cadaver-based values represent a "relaxed" human, this is a conservative estimate of injury risk….'. The other sections do not mention any limitations. Each summary section should likely briefly comment on any limitations.

Response: An isolated discussion of limitations is included in each of the body region sections: neck, chest, abdomen, knee/thigh/hip, and lower extremity, as well as the discussion of the fleet and field estimated injury risk comparison chapter. The limitations of the head injury risk function development are included throughout the chapter, as well as in the referenced literature.

R3: The regulatory intent with BrIC is nebulous. It is explicitly stated that HIC15 will be used as a head injury risk metric. If BrIC will be in the proposed regulation that needs to be stated just as explicitly. Section 3 implies that BrIC will be used in the last sentence – I interpret that to mean for regulations, but that should be made clear.

Response: The report is not intended to describe application of the THOR-50M in regulation or consumer information testing; rather, the intent is to document the methods and results for deriving injury criteria that can be applied to the THOR-50M ATD for the benefit of researchers, automobile manufacturers, restraint system developers, government agencies, consumer advocates, or any other institution conducting tests with the THOR-50M ATD. Several instances of the phrase "NHTSA intends to use" in the Summary have been removed or revised to avoid confusion.

Introduction

R1: Although it is cited, the Parent 2016 document is publically unavailable.

Response: References to the THOR Biofidelity Report have been replaced by the Parent et al., 2017 Stapp paper, “Biofidelity Evaluation of the THOR and Hybrid III 50th Percentile Male Frontal Impact Anthropomorphic Test Devices.”

R3: I would add a paragraph that describes what the agency intends to do with these injury criteria. Is everything in here going to go into regulation? Or are some of these criteria intended for research and product development rather than regulation. Or perhaps this a presentation of the state of the art IARVs with no decisions made yet on what will go into the future regulation.

Response: This report simply presents a set of potential criteria and the methods used to develop them. It is not the intent of this report to detail potential consumer metric/regulatory applications.

R3: Table 1.1 – The table title and/or column headers should clearly indicate if the totals are annual, or over the 5 years of NASS-CDS data. I assume these are counts for all 5 years based on reading the text. In my opinion, the table would be most informative if all the numbers were annual values. Same goes for Table 1.2

Response: Text has been added to state that the table is presenting a total over the years included.

R3: Table 1.3 – I assume this is total cost per victim over the course of his/her life (not annual costs and not costs over the NASS-CDS case years 2010-2014), but that should be made clear.

Response: That assumption is correct. The text has been modified to make this more clear.

R3: Figure 1.2 – again, is this annual costs or over the full 5 years of NASS-CDS cases? This would be better delivered as annual costs.

Response: Text has been added to state that this is a total cost over the years included.

R4: *Introduction, Section 1.1.* Clarify that FARS only provides information on fatalities on public roads.

Response: Change has been made to note that the FARS data applies to public roadways.

R5: Table 1.1 & 1.2: I am not sure what the column labels mean (pt, lwr, upr)

Response: Text has been included to describe the what pt, lwr and upr represent.

Methodology

R2: Statistical Assessment / Creation of Risk Functions (pg 16) – For your three main measures of model fit/predictability, using the area under the receiver operating characteristic curve is a good performer when using left censored or right censored data. If the datasets you are referencing are heavily interval censored (having a non-injury and injury point for a single PMHS specimen), you may be able to improve your predictability using a Brier score as your first discriminator. (Layana, 2016)[1]. And in addition to your Hosmer-Lemeshow's goodness of fit test, you may consider reviewing the Kolmogrov-Smirnov to supplement your selection of the most predictive biomechanics parameter (Yoganandan, 2016)[2]

> *Response*: The vast majority of the data used in this report was left/right censored and not interval censored. Thus, the methods presented are appropriate.
>
> The use of the Komogrov-Smirnov test as used in Yoganadan et al. 2016 does not appear to have a described benefit beyond what is accomplished by the Hosmer-Lemeshow goodness of fit test and has not been used at this time.

R1: Section 2.5 states that 'Extrapolating forward to the potential time frame in which THOR is planned for application and the estimated mean age for male drivers in frontal crashes is 40-years old.' However, the evidence to support this, Figure 2.1, does not show the extrapolation to the time frame in which THOR will be planned for use. Additionally, no comments are made as to the anticipation of how this trend might change with the aging population.

> *Response*: Language changed to simply extrapolate to 2020. The resulting age of 40 could be considered a placeholder at this point. Applications of the criteria presented herein that include age as a covariate can consider use of other ages as needed depending on the intended target population.

R1: Have the authors considered showing the age distribution of the driving population as opposed to the mean? This might allow subsequent scaling of injury criteria based on select age bins which would then quantify the range of age-based injury risk.

> *Response*: The average or mean driver age was chosen simply for having a nominal age to use for risk functions that include age as a variable. For any criteria w/ age as a covariate, the end-user can apply different ages if they desire to evaluate risk for an age different than the average of 40 suggested in this report. The limitation would be to not use ages outside the age range of PMHS used to derive the subject criterion.

R3: Figure 2.1 – You need to present the data as a scatterplot – the spline implies a continuous data set.

> *Response*: Change made.

---

[1] Lavanya, A., and T. Leo Alexander. "Confidence Intervals Estimation for ROC Curve, AUC and Brier Score under the Constant Shape Bi-Weibull Distribution." *International Journal of Science and Research* 5.8 (2016): 371-378.

[2] Yoganandan, Narayan, et al. "Deriving injury risk curves using survival analysis from biomechanical experiments." *Journal of biomechanics* 49.14 (2016): 3260-3267.

Head

R1: The structure and style of the 'Head' section is the only anatomically-focused section that is inconsistent with all others. It should be modified to both be consistent and to include the necessary sections that were omitted. This would include 'Design', 'Data' and 'Limitations'. It seems that the topics are already partially addressed in the text, but each of these should be handled explicitly by inclusion as a dedicated section.

Response: The requested change is not necessary as the Head section is different than other sections in terms of the information presented.

R1: GHBMC and SIMon introduced for the first time on page 27 without prior acronym definition.

Response: Acronyms have been spelled out with first use.

R1: Brain Injury: The use of BRIC is explained clearly and justified by the peer-reviewed works cited within the text.

Response: Thank you for the thorough review of the brain injury section.

R1: Brain Injury: While the inclusion of additional data (Section 3.4.2) to test the correlation between BrIC and CSDM/MPS strengthens confidence that the relationship is robust, it is not clear as to how these ATD data (BrIC) were compared to CSDM/MPS. Presumably the GHBMC or SIMon model were used…or both?

Response: BrIC is derived from CSDM or MPS, which themselves were developed through simulation using primarily SIMon as described in Takhounts et al. (2013). The correlation figures shown in Section 3.4 simply show the correlation of CSDM and MPS vs. BrIC for the respective ATD tests. BrIC in these ATD tests is calculated directly from the angular rate data, while CSDM and MPS is from SIMon simulations utilized the kinematic time-history data of the ATD headform. A sentence was added to the text to clarify this process.

R1: Brain Injury: Discussion of the Pareto Principle seems unnecessary.

Response: Section deleted

R1: Brain Injury: BrIC Summary is mis-numbered as 3.2.8

Response: Renumbered in a correct order.

R1: Brain Injury: The statement was made about the '"stability" of the BrIC versus CSDM relationship irrespective of the ATD and the test mode'. It would be good to comment on the fact that the various ATDs respond to similar test conditions with varying magnitudes, but that the general trend of the relationship with CSDM does not venture from the trend line.

Response: The authors believe this this is described sufficiently in Section 3.4.3, as well as in Takhounts et al. (2013). As this report focuses on the THOR-50M, expanding on the existing discussion of other-than-THOR ATDs is not warranted in the report.

R1: Skull/Facial Injuries: Use of HIC15 is justified based on its correlation to skull fracture and its historical basis.

Response: Thank you for the thorough review of the skull/facial injuries section.

R4: *Section 3.4.5, Pareto Principle.* I suggest simplifying or eliminating this section as the case that “New BrIC” is not a meaningful improvement over the original BrIC is made very effectively in the preceding subsections (i.e., new BrIC only has a slightly better fit to CSDM and AUROC for new BrIC and original BrIC are effectively identical).

Response: Removed.

R4: *Section 3.4.7 and Summary.* More detail about the work described in the Appendix is needed in section 3.4.7. I would suggest including a sufficient level of detail so that the reader can understand the rationale for including the findings from the DOE and optimization studies in the summary. I would also suggest simplifying the parts of the summary relating to the optimization study and referring to the Appendix for additional detail.

Response: Thank you for the suggestion. The DOE and optimization study information previously included in the summary was moved to Section 3.4.6, and the summary was revised accordingly.

R4: *Summary*. The section numbering on the summary is incorrect.

Response: Corrected

R4: Section 3.6. Some explanation of the figures in this section is needed as is text to refer the reader to the appendix for ATD test data.

Response: A discussion of the figures in Section 3.6 was added to the report, as well as in similar sections in other body region chapters. The details of the data included, along with a reference to the appendix with more information on the included tests, are described in Section 2.4, as this information applies to all body region chapters.

R4: Figures 3.7, 3.8. 3.9., 3.20, and 3.21 are difficult to read (look pixelated). Suggest including higher quality images in the final version of this report.

Response: Where possible, the publication quality of these figures has been improved.

R5: BrIC - Ultimately, BrIC is a transfer function relating CSDM / MPS to a rotational velocity metric. This could be helpful to not have to run a brain FE model; however, the approach presented for developing BrIC is an extrapolation of the data and may not be valid.
As I see it, there are a few areas of concern here. 1) the relationship between the CSDM / MPS metrics and injury (and their large uncertainty), 2) scaling animal data to humans can be difficult 3) adequate model validation of the FE brain models to accurately predict brain displacement and pressure and thus CSDM, 4) the approach for developing injury risk functions for each AIS level, and 5) the correlation between strain and BrIC.

1. The injury risk functions for both CSDM and MPS have significant confidence intervals (CI) at low values for severe injuries. For CSDM, the CI at CSDM=0 is ~0.2 and ~0.5 for SIMon and GHBMC respectively (Figure 3.8). MPS is slightly better, but only for AIS4 injuries (see below for comments on extrapolating to other AIS levels).
    a. Figure 3.8: in a) and b), the error bars on the fit are large, I don't know that the curve has much meaning especially when trying to extrapolate to lower AIS levels.
2. It was not clear how scaling animal data fit into the overall approach for developing BrIC. I assume it is related to the strain correlation to injury; however, without sufficiently validated models of both the animal brain and human brain linking human brain motion to strain-inducing injury in

animals carries much uncertainty. Even using scale factors as suggested here doesn't account for the differences in brain geometry.

3. For SIMon, going back to Takhounts 08, based on an eyeball-CORA of the data, the model does not predicting either pressure or displacement with any confidence, thus calling into question the strain measure outputs. It also isn't clear from any of the documentation what types of impacts the models are validated for (e.g. direction, magnitude, energy, etc.). For example, the Nahum dataset only has linear acceleration, so even if SIMon matches the pressures of the testing, it isn't necessarily validated for rotational energies. The data in Takhounts 08, does not indicate the range of dynamics of input loads used in the correlation. For GHBMC, I was unable to get a copy of Mao so I don't have a good grasp of how well that model reproduces pressure and displacement.
4. I do not believe this extrapolation to lower AIS levels is valid for use as a regulatory metric. For AIS≥1, the injury risk function is nearly vertical and starts at a CSDM of nearly zero. Given this info, a CSDM of ~0.05 would have a 100% probability of an AIS≥1 brain injury (concussion). If the confidence intervals were included in the lower AIS level extrapolation, they would be nearly meaningless metrics.
5. Figures 3.7, 3.10 & 3.16: both have intercepts around 0.5, implying that the BrIC metric is not valid for values below 0.5, since anything below 0.5 corresponds to a CSDM of 0. Ultimately, BrIC should be directly related back to the underlying injuries using matched pair testing. Using a model and extrapolating the data results in a risk metric with little confidence. Also, the data used in developing the BrIC metric should be documented in the appendix like all of the other metrics.

Ultimately, although a head rotational metric is needed, I do not agree that BrIC is appropriate for inclusion in future regulations with further research.

Response:
1. We agree that confidence intervals are wide, but so are responses of humans/animals to the external stimuli. Reducing the width of the confidence intervals can only be done by collecting additional injury data, otherwise manipulation of confidence intervals (presumably through model overfitting) will only lead to incorrect interpretation of the variability of responses in the human/animal population.

2. The approach to scaling animal data in the development of SIMon is addressed in detail in Takhounts et al., 2003. In summary, two approaches were attempted: scaling the SIMon model to the size of the animal specimen and applying the experimentally-measured kinematics, and scaling the experimentally-measured kinematics and applying to the unscaled SIMon model. Ultimately, the latter was selected. This process fits into the overall strategy of developing BrIC as described in Section 3.4.1, specifically the second paragraph. Scaling of animal data to humans does carry significant uncertainty. However, unless live human subjects are tested in the same environment as the animals, this inherent uncertainty is impossible to overcome.

3. Both GHBMC and SIMon models were validated to the same set of data. As discussed in detail in Takhounts et al., 2008, there are indeed limitations in the validation of the SIMon model, including (but not limited to) the different sizes and ages of the subjects, testing at different labs using different measurement techniques, and the accuracy and sensitivity of the measurement techniques themselves. Additionally, the limitations of the experimental strain data measurement using neutral density targets was discussed in detail in Drake et al., 2017[3]. Given these limitations, it would not be

[3] Drake, A. M., Takhounts, E. G., & Hasija, V. (2017). Investigation of parameters affecting brain model validation and brain strains using the SIMon finite element head model. In Proc. IRCOBI Conference (pp. 410-431).

possible for a single model to replicate each experimental measure. Regarding validation under "rotational energies", while the validation conditions were conducted using linear impacts, these impacts induce rotational head motion, else there would be no NDT displacements or strains. Thus, the models were indeed validated based on rotational head kinematics.

4. Please refer to Takhounts et al. (2013), in which other data from college football was used to validate the AIS 2+ curve derived from scaling. Due to lack of data on AIS 1+ brain injuries, the given curve gives the best estimate of the injury risk at that level. As the commenter rightfully suggests, extrapolation of the risk functions to a level for which no validation data exist would result in a risk function of questionable meaning with exceedingly large confidence intervals. As such, the end user should consider the relevant limitations if applying the AIS 1+ risk function.

5. Regarding the relationship between BrIC and CSDM, the Y-intercept is simply an artefact of the selected correlation function, as the CSDM is effectively asymptotic to the Y-axis. A logarithmic function could be used instead, but that would add unnecessary complexity for questionable value. A BrIC value of below 0.5 (0.523, to be precise) would correspond to a CSDM of 0, which in turn would correspond to an injury probability of 0.

Regarding matched pair testing, this would indeed be ideal, but the inherent limitation of brain injury risk function development is that the metrics used to assess brain injury (CSDM, MPS) cannot be measured *in vivo.* Instead, animal tests and human volunteer tests have been used to validate both modeling techniques (SIMon, as described in Takhounts et al. 2008) and injury criteria (BrIC, as described in Takhounts et al. 2013).

The data used in development of BrIC is listed in Appendix A of the report, and is publicly available through the NHTSA Vehicle Crash Test Database (https://www-nrd.nhtsa.dot.gov/database/veh/veh.htm) or the IIHS TechData site (https://techdata.iihs.org/).

R5: HIC15 - Similar to my comment above about BrIC, the underlying data for the HIC injury risk functions should be shown in the Appendix. Instead of using the existing HIC curves, has anyone attempted to recalculate the functions using existing and new data (e.g. football concussion data) using the newest techniques (e.g. survival method)? This could give more sensitivity to AIS≥1 and AIS≥2 level injuries at low probabilities. Another concern is related to the validation space for this metric. Is HIC in its current form valid for oblique impacts at the magnitudes and ΔV. It is critical to define this range, so that in the future, others will not misuse the injury metric.

Response: HIC has a long history of use and acceptance for, among other things, side impact, frontal and frontal oblique conditions including both belted and unbelted conditions for the frontal crash conditions found in FMVSS 208. While using new data and creating new risk functions could be considered, we felt it out of scope for this exercise given there is nothing unique about the THOR-50M in terms of the biofidelity, instrumentation and/or associated response of the THOR headform under hard contact conditions as compared to other ATDs already in NCAP/regulation.

Regarding the validation space of this metric, this topic is addressed in response to a similar comment in the "General Comments" section.

R6: We fully support NHTSA's decision to include a rotational-based brain injury criterion for crashworthiness assessment using THOR and we compliment NHTSA's continued development of BrIC. In particular, we are pleased to see that additional crash test data have been used to assess the strength of the BrIC-CSDM correlation, and that the agency has decided to adopt the CSDM- based BrIC injury risk functions (IRF). As you are aware, UVa and others have shown which suggested that the original MPS-based IRFs in Takhounts et al. (2013) over-predict brain injury risk. Along these lines, we believe there are remaining aspects of BrIC that should be further improved to strengthen the brain injury risk assessment.

Below is a list of specific comments:

1. We commend the agency's continued effort to assess the effect of head impact duration on brain deformation and the associated correlation between BrIC and CSDM. We have concerns, however, that this analysis is insufficient in and of itself to justify the exclusion of impact duration and/or angular acceleration in the mathematical equation of BrIC. NHTSA and others have established a sensitivity of critical velocity values to duration. While NHTSA has done an analysis of sensitivity and chosen values based on the current data set (most frequent duration), this duration does not necessarily relate to future test conditions and countermeasures. In fact, the parametric evaluations of duration demonstrate that there can be significant differences depending on which duration regime is considered. Thus, we propose the use of either time-dependent critical values and/or the addition of angular acceleration. For example, the addition of angular acceleration has been shown to improve prediction of CSDM across all crash test modes (Gabler et al. 2015[4], Gabler et al. 2016[5]).

Response: We disagree with the statement that angular acceleration improves prediction of CSDM. In the referenced papers, angular acceleration was filtered at the CFC60 (instead of CFC1000 that is recommended by SAE J211 for a filtering class for accelerations), which in effect "forces" peak angular accelerations to correlate with peak angular velocities. This correlation (collinearity) between the two parameters may slightly improve overall correlation to CSDM, especially when several additional terms are added to already existing formulation of BrIC. For more details on the effects of angular acceleration on CSDM, filtering frequencies, and collinearity leading to overfitting see Kelkar et al. (2020)[6].

2. Recognizing that brain deformation depends on the duration of head impact in addition to just peak angular velocity, the agency has chosen to pursue a technique that involves direct calculation of head impact duration from an arbitrary angular velocity time signal (Section 3.4.4). However, we have concerns with this general approach to determine duration. Crash tests result in complicated head motions that consist of different pulse shapes and multiple local maxima due to vehicle and restraint interactions. Approximating these complex angular head motions with an idealized pulse may result in severe under- or over-estimation of brain strain due to differences between the arbitrary and idealized pulse shapes. While in general this has been addressed by NHTSA in terms of the overall regression, there can be significant deviations for any particular case of crash and countermeasure conditions. Finally, the process the agency has choose to determine duration of a signal (Appendix B) consists of

[4] Reference omitted

[5] Gabler, LF, Rodenberger, EJ, Crandall, JR, Panzer, MB. (2016) Predicting Brain Injury Using Head Kinematics. National Neurotrauma Symposium, Lexington, KY.

[6] Kelkar, R., Hasija, V., Takhounts, E. (2020). Effect of Angular Acceleration on Brain Injury Metric. In Press: expected publication at IRCOBI 2020.

numerous steps and we would request that NHTSA provide further clarification of the steps and procedure.

Response: The shape of the pulse for the time duration study was taken to represent most frequently occurring angular velocity shape experienced by the occupant's head in a car crash. As the reviewer mentioned, the overall correlation of BrIC to CSDM addresses the point above, e.g. if the signal shape was an important parameter and (as is shown in Appendix B) the time duration is important, then it would be impossible to correlate BrIC to CSDM in the way shown in Figures 3.7 a) and 3.9. If "any particular case" is the subject of interest then a correlate should not be used, rather an FE model along with CSDM should be utilized. A correlate, such as BrIC, by definition represents a population, not "any particular case". The same argument is applicable to the ATD design, which can not (by definition) represent any particular individual. Hence comparison of the ATD based measurements to any individual PMHS test is futile.

3. Although the agency has made modifications to BrIC to include duration effects ("New BrIC" and "Liberal BrIC"), these metrics do not correlate as well as formulations that explicitly incorporate angular acceleration (e.g., UBrIC formation – Gabler et al. 2016[7]) based on results of the GHBMC M50 brain model with a wide range of crash modes (Figure 1). While UBrIC requires three addition parameters for calculation determined by differentiating the angular velocity time signals, no additional sensors are required in ATDs. With proper sequence of differentiation and filtering, we feel this is a more robust method than approximating duration, and is a better predictor of the resulting brain deformation. We recommend that angular velocity data be filtered at CFC 60 (proposed for BrIC) prior to differentiating, as this has shown negligible effect on brain strain for angular acceleration frequencies above 100Hz (CFC 60).

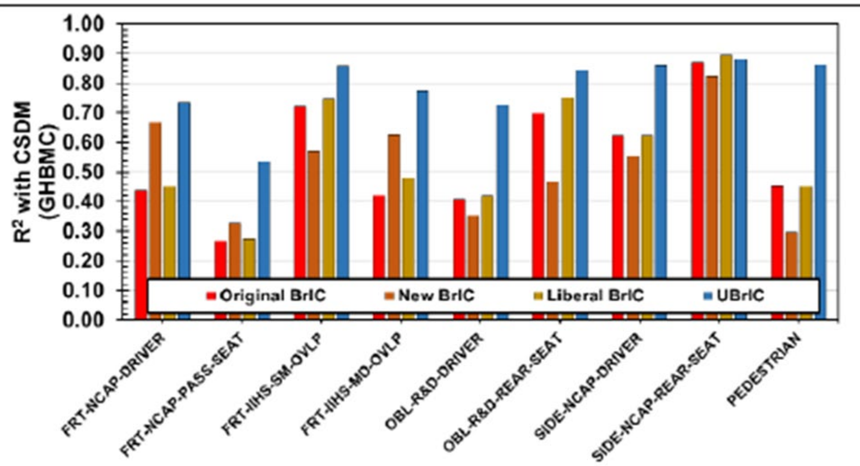


Figure 1. Correlation of different BrIC metrics and the UVa UBrIC metric by automotive crash mode and occupant location based on CSDM (25%) determined by the GHBMC M50 model.

Response: We have tried various differentiating methods (forward, backward, and central difference) and sequences (per Gabler et al., 2016) and found no difference at all between various differentiation methods and/or whether the signal was differentiated prior or after filtering. The last sentence of the comment is important because it demonstrates that peak angular acceleration has no effect on CSDM for frequencies above 100 Hz. According to Gabler et al. (2016): "Analysis of

[7] Gabler, LF, Rodenberger, EJ, Crandall, JR, Panzer, MB. (2016) Predicting Brain Injury Using Head Kinematics. National Neurotrauma Symposium, Lexington, KY.

angular kinematics revealed that angular velocity was sufficient to predict brain injury risk for impact durations up to 100 ms…" More details and analyses on the effects of angular acceleration on CSDM are given in Kelkar et al. (2020).

4. Based on the range of pulse durations reported in Figure 3.17 relative to those from Hardy et al. 2007, we recommend that FE models used for the development of BrIC be validated for longer duration head motions. Angular velocity pulses from Hardy et al., 2007 used for validation of the FE model span roughly 20ms to 80ms, depending on anatomical direction. Given that there are large number of occupant crash tests resulting in head kinematics with angular velocity pulses greater than 80ms, it is not clear that simulated brain deformations for these cases will be accurate. Thus evaluating the FE model's biofidelity under longer duration loading would strengthen the reliability of the FE models and their relationship with kinematic parameters.

Response: In Kelkar et al. (2020) we demonstrated that CSDM is negligible for time durations over 100 ms compared to those below 100 ms, hence validating the FE models for such time durations may not be beneficial. In addition, the vast majority of the time durations in car crashes are similar to those given in Hardy et al. (2007).

5. Throughout section 3.4 the relative performance of BrIC and its modified versions (New BrIC, Liberal BrIC) are assessed based on correlations with CSDM using all 749 tests. We believe it would be more useful if correlations were provided for each crash mode (e.g., Figure 3 above). For example, BrIC could have high correlation with frontal crash tests which may dominate the full dataset, but low correlation with oblique and side impacts. Furthermore, investigating correlations within each crash mode would help to verify the robustness of approximating signal duration with an idealized pulse shape, given that some crash modes are typically longer in duration than others. Finally, it would be useful to see the relative performance of Original, New, and Liberal BrIC for each crash test mode (as shown in Figure 1).

Response: Splitting the data into different crash modes leads to a small sample size for each mode, hence the correlation coefficient for BrIC naturally decreases due to small sample size. Once a significant sample size is accumulated for each mode this correlation will be assessed.

6. In Takhounts et al., 2013 a large dataset of direct head impact (pendulum tests) was used in the development of the BrIC-to- (CSDM and MPS) correlations. We are unclear if these pendulum tests included in the 749 dataset and the duration frequency histograms in Figure 3.17. If they are included they may be skewing the impact duration distributions given that these tests involve direct head impacts that are typically shorter in duration relative to crash tests. If the pendulum data were excluded, this would likely increase the median duration, and thus the critical values associated with Liberal BrIC. In general, the issue related to what data sets are included and what are [not?]

Response: As was described in Takhounts et al. (2013) pendulum impacts were of various time durations and intensities (padded, unpadded, at various impact speeds, etc.). These time durations are similar to those seen by the occupants in car crashes and hence are relevant to the proposed application in NCAP tests.

7. Appendix B – Step 3b: "If the peak is too close…" Is goal to get the duration to be symmetric about the peak (is the duration always supposed to be symmetric about the peak)? If so, then why is step 3,b,(i & ii),2 needed, since 3,b,(i & ii),1 is already a symmetric calculation? Step 4: "Counter is stopped when the first maximum in area is found in the first 100ms". Does this require that there be at

least one maximum in area before 100ms, and that the duration calculation should always start no later than 100ms?

Response: No, the duration does not have to be symmetric about the peak. The answer to the last two questions is – yes. The routine allows for calculation of the peak within the first 100 ms.

8. The discussion of the Pareto effect in Section 3.4.5 seems inconsistent with the level of rigor and documentation provided other sections of the document. While the Pareto Effect can obviously be identified in numerous examples across various fields, none of the examples are relevant to the prediction of brain injury. Rather than the generic explanation of the Pareto Effect, we would recommend that NHTSA use a statistical measures that assess the benefits of additional predictors if they do indeed have concerns over the number of predictors.

Response: Removed.

9. Given the dependence of the injury risk functions on the underlying animal data sets used to establish the tissue based predictors, we recommend that the underlying animal injury dataset be made publically available and scrutinized by an independent entity prior to finalization of the brain IRF.

Response: Most animal data was derived from the publicly available data and may present a copyright issue. This is being worked on so that all the data is released and rigorously scrutinized.

Neck

R1: Regarding biofidelity, the following statement was made: 'In the frontal and lateral flexion conditions….the THOR-50M demonstrated acceptable kinematic and dynamic biofidelity…'. 'Acceptable' should be defined. Is it consistent with the criteria defined for biofidelity in the Introduction?

> *Response*: This was changed to refer to the Parent et al. 2017 paper regarding the THOR-50M biofidelity, which also includes definitions of "good" biofidelity.

R1: Section 4.6.1 states 'A limitation to this test series is that PMHS neck loads were not directly measured. Due to this limitation, the matched pair tests are used herein to relate PMHS outcome to THOR measurements in the same test condition.' This seems to assume that the peak ATD forces equate with the time of injury. This is not necessarily the case. For example, if the specimen sustains an injury, it is presumed that forces, if measured, would likely decrease. However, the ATD forces could continue to increase beyond the time of cadaver injury since the ATD is not frangible. Additionally, if the neck is injured due to a combination of loading conditions, then isolated (peak) forces are not alone a good predictor of the injury/response. Due to the combinatorial effect, an injury can be sustained at non-peak forces. Therefore, utility of these particular whole body PMHS-ATD comparisons seems limited beyond simply observing outcomes and injury results.

> *Response*: While this is a valid point, only three of the whole body PMHS-ATD tests resulted in AIS 3+ injury, and these would be the only three datapoints affected if the ATD forces did increase beyond the time of cadaver injury (the 45 non-injured datapoints are not affected).

R1: The US Army WIAMan Project has generated an extensive amount of well-defined neck compression data that may be of use to address the noted gap: 'there is insufficient information to develop a transformation for the THOR-50M response'.

> *Response*: Without THOR-50M matched pair data, the additional data referred to by the reviewer cannot be used to develop transfer functions.

R2: Legend for Figure 4.2 (pg 56): Recommend adding a second digit to your vertical axis in figure 4.2, "Neck injury rate by model year (1993 to 2012) for frontal crashes in NASS-CDS 1993 to 2014." The current axis has 0.0%, 0.1%, 0.1%, 0.2%, 0.2%, and 0.3%. I'm assuming this is a rounding issue and should be 0.00%, 0.05%, 0.10%, 0.15%, 0.20%, and 0.25%.

> *Response*: Figure 4.2 was revised to include 2015 CDS data, and in the process the vertical axis labeling issue was resolved.

R3: … the report is not clear on which sensors will be used to assess Nij – it should be explicit in this regard.

> *Response*: Section 4.10, "Application of Risk Functions to THOR-50M," states that the force and moment are taken from the upper neck load cell.

R3: … there has been almost no biofidelity assessment in the most important mode of loading for the neck – compression. Assumptions are made about this in lieu of simple assessment.

> *Response*: Biofidelity assessment of the THOR-50M in compression was not carried out because, in frontal and frontal oblique crash tests, the axial response of the neck is dominated by tension.

R3: Finally, there is an error in the transformation of stiffness to load in the formulation of the IARVs, which results in values that I believe are too large by more than 40%. I will explain my reasoning in below and in detail in the attached PowerPoint slides.

Response: The PMHS-THOR transfer function was adjusted to account for equivalent energy at failure.

R3: Page 55: The "entire neck structure" can experience compression, while the location of injury can be in a state of combined compression and bending. Therefore, the sentence should read, "That is, the chosen mechanism represents the type of loading/motion experienced at the site of injury."

Response: The mechanisms attributed in CIREN are not necessary specific to the site of injury (they are regional mechanisms, with the neck being a region).

R3: Figure 4.2: It's hard to imagine a 0% injury rate in 1990. Also, it looks like the regression was forced through 0. Was the "trend" statistically significant? If so, that should be stated.

Response: Figure was revised to include 2015 CDS data and calculated as a three-year running average; the injury rate no longer intersects 0% in 1990.

R3: Section 4.4 Instrumentation – I gather from the Summary and the context of this section that although the cable forces are measured, they are not used in the biofidelity assessment and in the IARVs. This should be stated explicitly since there is still some confusion about this among the OEMs and academies. If the lower neck load cell readings are not in the regulations, will the sensor be optional?

Response: Application of the risk function to THOR-50M is shown in Section 4.10, which specifies that the forces from the upper neck load cell are used in the Nij calculation. As described earlier, this document does not define requirements for regulation or consumer information testing.

R3: Section 4.5 Biofidelity – Some of the data presented is from our report (authored by Jason Luck, 2014). I would like to see that report referenced so that Dr. Luck gets appropriate credit for this work and so the reader can find the source. The same goes for some of the methodology in Section 4.6.

Response: This reference has been added.

R3: Page 60: Please explicitly reference Dibb in section 4.5.1 as well as Luck for the THOR- 50M data.

Response: This reference has been added.

R3: Figure 4.6 – please reference the Luck report

Response: Added "adapted from Luck et al. 2014"

R3: Section 4.5.3 Compression - There is a lot of data on the human in compression and almost none on the THOR. This is a significant limitation and one that is easily remedied. Neck injury is rare, but it is still a huge expense (Table 1.3). Since most catastrophic neck injuries are from compression, I think the agency leaves itself open to criticism if it does not take the obvious step of funding some research on this. In my opinion, it is a glaring omission given how much work has gone into the other modes of loading.

Response: This has been described as a limitation of the approach. It is also noted that for in-position occupants, THOR-50M should rarely experience a peak Nij in either compression mode. Only one of the THOR-50M fleet tests examined had a peak Nij in compression (Nij=0.3), while the remaining tests all had the peak Nij in either tension-flexion (28%) or tension-extension (69%).

R3: Section 4.6 Data: None of the Tables in Appendix F are referenced. Please make specific references to those table in this section and elsewhere where appropriate. If there are published journal articles or reports on these data, please reference those as well.

Response: References to the tables in Appendix F have been incorporated.

R3: Page 63: If there a reference for the fleet tests referenced in the second sentence? Either a report or an appendix listing should be included.

Response: A table in the appendix has been added.

R3: On the 54,000 N/s loading rate. In my opinion, this is significantly overestimated. See my comments below (Page 64) for more details. Because the THOR is stiffer than the human in tension, it will give a higher loading rate for a given crash. The loading rate is dictated by the natural frequency of the head (the mass) and neck (the spring). Since the THOR head mass is the same as the average human, the difference in loading rate is therefore driven by the square root of stiffness. Assuming things are linear, and the THOR is twice as stiff in tension as the human, you should divide the THOR rate by root 2 to get the equivalent human loading rate. By my calculation, that should be around 38,200 N/s.

Response: In response to this comment, the PMHS-equivalent load rate was adjusted to 38,200 N/s.

R3: A reference for the UVA PMHS tests should also be included. At least, reference Table F.2.

Response: References have been added.

R3: Page 64: Peak force is not linearly related to stiffness for energetically equivalent events. It is related to the square root of stiffness. Therefore, the scaling factors are too large. In my opinion, the factor of 2 for compression and tension, which is used to scale the human IARV to the dummy, should be 1.41. Likewise, the factor of 1.4-1.3 for flx/ext should be 1.16. The reason for this is explained in the attached PowerPoint. Clearly, this has implications throughout the rest of the section and also for the other body regions if a similar transformation method was used.

Response: The PMHS-THOR transfer function was adjusted to account for equivalent energy at failure.

R3: Section 4.6.2 – This is a nice approach to rate scaling the moment data – I wish I'd thought of it. My only critique is that it assumes that all the moment is reacted as the extreme limits of the geometry – in other words, in the supra-spinous ligaments and in the ALL. In reality, it is distributed in some way across the entire distance. In my opinion, a better estimate would come from using 1/2 the moment arm. It is possible that an even more accurate assessment could come from the ligament cross-sectional areas used by Cronin et al in the latest version of GHBMC model. I would also double check the 2/3 assumption on the location of the COR against the paper from Chancey et al. 2007.

Response: The center of rotation was adjusted to be consistent with Chancey et al. 2007. A 2/3 factor was used to account for the fact that not all of the moment acts at the extreme limits of the geometry.

R4: For the neck, I suggest basing the transfer function from dummy to human for tension on force at equivalent energy transferred to the c-spine structure rather than force at equivalent deflection. Similarly, for rotation, I suggest basing the transfer function on moment at equivalent rotational energy rather than moment at equal rotation. The rationale for these recommendations is that in the field, energy transferred to the dummy and the human structures is more likely to be equivalent for a given crash severity than either applied force or deflection. Matching force at equivalent energy transferred to the c-spine structure also accounts for the fact that the PMHS fails, while the ATD does not, in matched pair tests. Note that these recommendations are based on unpublished work presented by Duke, which I believe that NHTSA is aware of.

> *Response*: The PMHS-THOR transfer function was adjusted to account for equivalent energy at failure.

R4: *Section 4.1, first paragraph*. Suggest editing the text “It is important to note that the mechanisms show are regional mechanisms, not organ-specific mechanisms” to clarify that the neck is not an organ.

> *Response*: This language was updated per the comment.

R4: *Section 4.5.1, Tension and Section 4.5.3, Compression.* I recommend comparing THOR and DAHNM responses at equivalent energy rather than equivalent deflection.

> *Response*: In response to the comment, the comparison has been made on an equivalent energy basis.

R4: *Section 4.5.2, Flexion/Extension.* Please clarify the regions over which □My/□Ma and □My/□□ calculated. An explanation of why a single slope is appropriate to describe the ATD curves in Figure 4.6 would be useful. I also recommend comparing ATD and DHANM model responses at equivalent energy rather than equivalent rotation as the underlying rotational stiffnesses are different.

> *Response*: The moment-angle comparison was updated such that it accounts for the region of approximately linear stiffness involving head displacement between 20 and 60 degrees in flexion and extension (rather than using a single point). The scaling factor was also updated for equivalent energy.

R4: *Section 4.9*. Suggest adding some text to explain the figures in this section and pointing the reader to the appendix for the associated data.

> *Response*: Done

R5: For Nij, why was the upper neck chosen for use as the metric? In the CIREN data reported, it didn't specify at what cervical level the injuries are occurring. I suspect that lower neck metrics may better predict seen injury patterns.

> *Response*: CIREN data demonstrates that about half of AIS 2+ cervical spine injuries occur in the upper and half in the lower cervical spine. Furthermore, recent analysis using human body model has demonstrated that tension in the upper neck actually correlates well to strain throughout the neck, giving some confidence that injury metrics for the upper neck can be predictive of all neck injuries.

R5: For all of the components, I am concerned that excluding females may skew the resulting vehicle designs. Although the THOR is intended to represent a 50th percentile male, in lieu of a female THOR, these injury metrics are the only control for mitigating injury for all occupants whether male or female. Since female neck strength is significantly lower than males, adopting male only injury risk functions results in injury metrics and ultimately regulations that are insufficient to protect all occupants.

> *Response*: All injury risk curves developed in this report are designed for males. Future work will evaluate small female injury risk, but that is outside the scope of the current effort.

R5: *Neck Tension* - On page 69, the Weibull equation doesn't result in a valid function. I think there is something missing or mistyped.

Response: Equation updated.

R5: *Neck Flexion and Extension* - Based on the failure loads for tension and the corresponding 50% injury risk value, using the mean failure rates may incorrectly characterize the actual risk. It is concerning that the extension moment is higher than the flexion moment, indicating this approach did not produce valid values. There are many cadaveric datasets out there for both flexion, and to a lesser extent, extension. I recommend conducting some limited matched pair testing with the THOR and developing better critical values for these Nij components.

Response: The data chosen for evaluating the critical intercepts were targeted to be "pure" loading. Many additional datasets are available but would not satisfy this constraint. It is not clear why the reviewer is concerned that the extension moment is higher than the flexion moment, since that is what the PMHS data showed. While the original Nij flexion intercept was greater than the extension intercept, those were inferred primarily from limited volunteer data and scaling. Other studies, such as the Carter 2002 study, showed that failures in "compression-flexion" occurred at much lower loads/bending moments than in "compression-extension", lending additional support for the critical intercept being higher in extension.

R5: *Neck Compression* - The minimal difference between AIS≥2 and AIS≥3 suggests that using mean failure rate may not accurately capture the 50% failure risk for each AIS level. In addition, the compression critical limit is higher than tension also suggesting the selected value is incorrect. Finally, compression not correlating well and being excluded from development of the Nij risk function indicates a strong need for additional research, and is recommended.

Response: Updates based on many comments have revised the risk curve, and compression data are not excluded. It should also be noted that none of the frontal sled tests and only 1 of the fleet tests have the maximum Nij in either compression-flexion or compression-extension. These modes are unlikely to occur with the expected use of THOR-50M.

R5: *Lateral Bending* - Since THOR is intended to be used in oblique loading, a lateral (Mx) and/or torsional (Mz) component of the Nij should also be included.

Response: It is unclear how lateral bending and/or torsion would be applied in an Nij formulation. It is possible that in the future, a separate injury criterion could be developed, if sufficient PMHS data became available.

Chest

R4: *Section 5.10.3, Survival Analysis*. It would be useful to describe censoring status for these tests. I assume that all non-injury tests were right censored and all injury tests were left censored.

> *Response*: The data presented in the peer-reviewed draft were initially treated as right-censored for non-injury observations and left-censored for injury observations, as described in the Methodology section. However, as another reviewer pointed out, several of the observations should be interval censored; though it makes a negligible difference in the outcome, the survival analysis was updated to include interval censoring for these observations. The text was updated accordingly.

R4: Section 5.11. Some text explaining the plots and referring to the appendix for the ATD data would be helpful.

> *Response*: A general description of the source of the fleet data is included in the Methodology section, and is not repeated in each body region section to avoid redundancy. However, we agree that some discussion of the plot or plots in each body region section would be appropriate.

R4: Section 5.12, Limitations. PMHS being unable to reproduce pulmonary injuries is a limitation that affects our ability to develop criteria for these injuries. However, in practice, this limitation is probably not meaningful as an injury criterion to limit rib fractures limits the mechanisms of pulmonary injury (i.e., chest compression and rate of compression).

> *Response*: The referenced limitation was intended to address the uncertainty of injury determination and its implications on the injury risk function. If it could be reliably known that a PMHS sustained hemothorax or major pneumothorax, that observation could be considered an injury observation even if less than three rib fractures were sustained. While limiting chest compression could also be protective of pulmonary injury assuming that the loading surfaces are blunt and do not create penetrating wounds, the recommended injury risk function could be different if reliable hemo/pneumothorax diagnoses were available. As can be shown from field data, there are instances of hemothorax or major pneumothorax occurring in absence of three or more rib fractures (for example, see CIREN CaseID 359831171).

R5: As mentioned on other metrics, is AIS≥3 the appropriate injury severity? Given vehicle designs today, a lower severity metric would be desirable.

> *Response*: Selection of an AIS level to target for injury risk reduction involves considerations for both the incidence and the severity of the injury, and largely depends on the application. Generally, since the incidence of MAIS 2+ and MAIS 3+ injuries follow the same trend (Figure 5.2) and the AIS 2+ and AIS 3+ injury risk functions have a similar shape (Figure 5.15), countermeasures intended to reduce the risk of AIS 3+ injuries would also reduce the risk of AIS 2+ injuries at a similar rate. An AIS 2+ injury risk function was added to the text for the benefit of users with applications specifically focusing on AIS 2+ injury.

R6: Overall, the Agency has come to a reasonable result in the development of the thoracic injury evaluation tools for THOR. The process has some flaws, however, and these should be addressed so that these problems do not undermine the end result. We believe the feedback below, if implemented, would strengthen the foundation of the proposed criterion and age- dependent injury risk function, but likely will not change them appreciably.

Below is a list of specific comments:

1. Figure 5.2: There seems to be something off with these data. This plot shows that the MAIS2+ and MAIS3+ chest injury rates drop to nearly zero for older model years around 1990. This is very consistent with the corresponding analyses for the other body regions as well – the Agency shows that the injury rates observed for all body regions except for the tibia/fibula drop to nearly zero around 1990, and the injury rates show an increasing trend with newer model years for all body regions except tib/fib, foot/ankle, and skull/face. Is it the Agency's contention that the injury rates for all of these body regions were nearly zero in 1990-1991 as indicated? It is difficult to assess whether or not this analysis is valid without knowing more about how it was conducted. We recommend verifying the analysis to make sure that there was not some methodological error forcing the injury rates in the early model years towards zero. We also recommend describing the methods used in this analysis in greater detail so that the reader can assess the validity and applicability of the methods and results.

Response: The plots have been updated to only include MY 1990+ or a running 3-year average for MY 1992 and forward. This change was made because very few belted drivers with deployed airbags existed in the data set prior to MY 1990. This same change was made to all body region graphs for injury vs. model year.

2. Table 5.2: As described in McMurry et al. 2016 (referenced in the text), the sled test dataset includes three PMHS that were subjected to repeated tests in two test conditions (the 40 km/h and 10 km/h Gold Standard tests of Lopez-Valdes 2010). As a result, these tests are not truly independent datapoints, and should not be treated as such in the final IRF regression model. McMurry et al. (2016) accounted for this by treating the results from those PMHS as interval censored in the survival analysis, with the results from the 10 km/h tests representing the lower bound of the censored interval, and the results of the 40km/h tests representing the upper bound. Were similar steps taken in the current analysis to account for the fact that these datapoints represent repeated measures? If not, appropriate steps should be taken to account for this in the analysis. (Note: It is not expected that this would change the resulting model substantially, but it would help improve the validity of the process used to develop the model.)

Response: The data presented in the peer-reviewed draft were initially treated as right-censored for non-injury observations and left-censored for injury observations, as described in the Methodology section. As suggested, the survival analysis has been updated to treat the repeated subjects as interval censored, and the text was updated accordingly. As expected, the influence on the resulting model was negligible.

3. Table 5.3: NHTSA added PMHS tests from a study with a 1986 Ford Tempo buck, with the PMHS seated in the driver position (3-pt belt+airbag, lap-belt+airbag). Because NHTSA didn't have THOR-M tests in this condition, the Agency matched those PMHS tests to THOR-M tests in a 1997 Ford Taurus buck with the PMHS in the passenger position. The notion of matched dummy-PMHS tests relies on having equivalent exposure conditions between the PMHS and dummy. In this case, the seatbelt, retractor, airbag, airbag type (passenger vs. driver airbag), seat, knee bolster, and the buck acceleration pulse are all different between the '97 Taurus dummy tests and the '86 Tempo PMHS tests. There is also a steering wheel present in the PMHS tests (performed in the driver position), while there is not a steering wheel present for the dummy tests (performed in the passenger position). Therefore, it cannot be reasonably assumed that there is equivalent exposure between the dummy and PMHS tests in this case. Figures 5.6, 5.7, and 5.9 suggest that adding these tests to the dataset has little-to-no effect on the resulting IRF. Since the addition of these data prompts substantial methodological questions with little-to-no benefit, we recommend removing these tests from the dataset.

Response: Agreed; the SledOnly data set is now used in development of the final model, and the text was updated accordingly.

4. Table 5.7: NHTSA selected the final dataset based on how well it fit the model. This approach is backwards, as deficiencies in model fit are not reflective of deficiencies of the data (assuming the data are accurate), but are reflective of deficiencies in the choice of model. NHTSA should select the dataset that they wish for the model to cover (in other words, the dataset covering the conditions for which they want to predict injury risk), and then find the model that most consistently fits the data. If NHTSA does wish to down-select the dataset, they should do so based on other reasoning.

One valid argument for down-selecting can be seen in Figures 5.6 and 5.7, which suggest that the relationship between chest compression and injury risk is different for restraint- type loading compared to hub loading. Because of this loading-type dependency, a single IRF should not be expected to be accurate for both restraint loading and hub loading.
Thus, either the loading type needs to be included in the model as a modifier, or the dataset should be stratified by loading type and distinct IRFs should be developed for restraint loading (as was done in the final recommended IRF) and hub loading. It may also be worthwhile to perform a statistical test (e.g., a log rank test) to determine if the difference between the impactor dataset and the sled-only data set is statistically significant. (Note that this will not affect the end result, as the final models included just the sled tests. This will just affect the reasoning behind down-selecting the final dataset to include only sled tests.)

Response: The intent of injury criteria development for the THOR-50M ATD was to enable prediction of injury risk using only the measurements collected by the ATD during a test. As such, requiring a parameter to define loading type in the prediction of injury risk was not desired. Additionally, based on the relative scarcity of data, an attempt was made to include as much of the available data as possible in the development of injury criteria. In the development of a thoracic injury criterion, the intent was to include all of the available data, including both sled and impactor observations. However, as shown in Figure 5.11, survival analysis using either the Sled+Impactor or ImpactorOnly data sets results in a much different shape to the injury risk curve which is much closer to a Weibull shape parameter of 1.0 than the remaining data sets. Further, the ImpactorOnly survival function shows divergent confidence intervals, which suggests the impactor tests alone would not be sufficient to generate an injury risk curve. Therefore, in addition to it not being desired for qualitative reasons, including loading type as a variable is not expected to result in an effective injury criterion. Unfortunately, this results in the exclusion of hub loading tests from the final dataset.

In the report, results of various datasets are presented for the benefit those who may have a different application of the injury criterion, though in response to this and similar comments, the rationale for selection of the final dataset has been updated to more clearly state the intent.

5. Following the above question, plots 5.6 and 5.7 indicate that the relationship between THOR chest deflection and injury risk is dependent on the method of loading application (hub vs. restraint loading). This suggests that the final recommended IRF should only be assumed to be accurate for restraint-type loading conditions consistent with those contained in the final (SledExtended) dataset. This represents a limitation of the IRF – dependency on loading type means that the IRF cannot be used as a standalone, independent means to assess injury risk. Instead, to confidently assess injury risk with this IRF, some information must be known about how the load is being applied (in this case, whether the chest deflection is a result of hub-type loading or restraint-type loading). Knowing that the IRF is dependent on loading method also calls in to question its utility for use with novel restraint types that

are not included in the current dataset – we don't know if the IRF would be the same or different for a completely different restraint type; given an observed dependency on loading method, we should not assume that the IRF would be consistent for other loading scenarios. The ideal solution is to develop an IRF that is not dependent on loading application, and can be applied to all cases without a priori knowledge of how the loading is applied. Was any attempt made to investigate alternative IRF formulations, seeking to develop an IRF that can be applied universally (for example, using information on the pattern of chest deflection from the four measurement locations to discriminate loading distribution)? At a minimum, given the observed sensitivity to loading type, the Agency should specify that the IRF should only be assumed to be valid for restraint-type loading conditions consistent with those contained in the final IRF dataset (SledExtended, or SledOnly if the 1986 Tempo tests are excluded as recommended).

Response: An attempt was made to investigate alternative IRF formulations during the process of developing the PCA methodology (Poplin et al, 2017). For the hub loading conditions, it was found that deflection patterns were not consistent with the correlation relationship observed in the restraint loading conditions. For example, the lower oblique hub tests resulted in lower deflection sum (LowTot) values disproportionately greater than would be expected based on the correlations seen in the baseline dataset when compared to the upper sum (UpTot), upper difference (UpDif), and lower difference (LowDif). Similarly, the frontal hub tests resulted in UpTot values that were disproportionately greater than would be expected based on the correlations with the LowTot, UpDif, and LowDif.

Since the deflections observed in the hub loading conditions result in fundamentally different chest deflection patterns, these conditions cannot be described by the same single reduced variable (i.e., component) that was developed using the principal component analysis with the original frontal sled test dataset. Also, given this implies different deflection patterns, it is also unknown if Cmax relates to chest injury risk in the same manner as it does in the baseline dataset. Because there is not a sufficient sample size in these test modes to allow for additional predictor variables (i.e., components) in the injury risk function to account for these differences, hub loading conditions are excluded from the final dataset.
Nonetheless, as discussed in an addition to the Limitations section, the even though the ImpactorOnly data sets were not used in the development of the injury risk function, injury predictions for observations in those data sets were relatively accurate.

6. Section 5.7, Predictor Variable, Stepwise Regression: NHTSA states that they used stepwise logistic regression, where the initial variable set included max deflection, deflection in all four measurement locations (i.e. four measurement variables), age, mass, stature, and sex. In other words, NHTSA had up to nine variables that they treated as independent in their initial models. There are two issues with this:

a. As indicated in McMurry et al. 2016 (also noted in Harrell 2001, Regression Modeling Strategies. Springer, New York, p. 60), there is a general rule when developing logistic regression models from censored data that there should be at least 20 datapoints (e.g., 10 injury and 10 non-injury) for each variable included in the model to avoid over-fitting the dataset. In other words, to fit a model with nine independent variables (without overfitting), one would need at least 180 datapoints. In its largest form, NHTSA's dataset contains at most 72 data points. When the dataset is stratified by loading type (due to the different relationship between deflection and risk observed for the impactor tests), this leaves 53 sled tests. If the 1986 Tempo tests are removed (as recommended above), this leaves 48 sled tests. Thus, the Agency does not have nearly enough data to fit a model with nine variables. This likely was a contributing factor to the

Agency's observation that mass, stature, and gender were not significant predictors of risk, as there is not enough data to independently assess these factors in addition to deflection and age. As a result, we recommend removing the statement that these factors were not found to be contributors as this cannot be accurately assessed in a dataset of this sample size, and as it is written it may mislead some to think that these factors would not contribute to the injury risk in the larger population. At most, given a viable sample size of 48 datapoints, any models investigated in this analysis should be confined to two independent variables. As age is likely to be a contributing factor (as demonstrated in this analysis and elsewhere in the literature), we recommend limiting this analysis to models that include age and one deflection variable as the two independent predictors in the model. (Note that this will not affect the end result, as the final models ended up constrained in this manner. This will just affect the preliminary analyses described in the document.)

Response: The agency did not attempt to fit a model with nine variables, not all at once at least. Stepwise regression was carried out three times. First to determine whether individual quadrant deflections should be included as predictor variables, second to determine whether the individual quadrant deflections explained any additional variance to the peak deflection terms, and finally to examine possible covariates. Additional discussion was added to the report to describe the first and second stepwise regressions.

During the first stepwise regression (using only the individual quadrant deflections), the largest model included three variables, but the resulting model retained only two variables. The second stepwise regression (which added the peak deflection terms) needed only one step, the entrance of the peak deflection term. At that point, none of the individual quadrant deflection terms were fit for entry into the model.

Once the decision was made to include only the peak deflection term, the final stepwise regression considered peak deflection, age, mass, stature, and gender. This stepwise regression included two steps, the entrance of peak deflection and then age, after which the remaining terms were not fit for entry into the model.

While the stepwise regression processes were primarily operating in forward selection, there were points where up to four variables were included in intermediate analyses. However, the fact that these variables were not ultimately included in the resulting stepwise model suggests that overfitting did not occur. Had all four or five variables been included, a variable reduction strategy would no doubt be desirable.

To avoid misleading readers, the phrase "for this sample" was added twice in the text to remind the reader that this analysis applies to the current sample, not necessarily the larger population.

b. In addition to the over-fitting issue described above, there is also an issue related to treating the five deflection measures (the maximum deflection and the deflection at each of the four measurement sites) as independent predictor variables in the stepwise regression. McMurry et al. 2016 demonstrated that in the sled testing dataset used here, the deflection measures from the four sites are highly correlated with each other. Since they are correlated with each other, if they are included as independent variables in a regression model then the model coefficients will be unstable (i.e., will have high standard errors). This is likely a contributing factor to the observation that adding deflection measures in addition to the maximum did not add information to the model risk prediction – since the deflection measures are all correlated in the sled testing dataset, a majority of the deflection information can be captured by a single measurement from each test. (Note an important distinction – this observation may not be true for other datasets where the deflection measurements are not correlated. This may have been a contributing factor to the observation that the risk prediction for impactor- type loading is different than restraint-type loading, if the deflection measurement correlation/pattern is different between those two loading types.)

Because of this, when considering a dataset with multiple predictors that may be correlated (e.g., similar measures taken at different locations), it is pertinent to check for correlation between the potential predictors before applying to a regression model. This was done for the sled testing dataset in McMurry et al. (2016) via Principal Component Analysis, which demonstrated that the deflections in the sled dataset could be adequately described by a single composite deflection measure. For the current document, constraining the analysis to a single, truly independent deflection predictor will not change the overall result, it will simply affect the preliminary stepwise regression analysis. To rectify this, we recommend that the Agency either add an analysis exploring the potential correlation of chest deflection measurements in the dataset (and discuss the implication of such correlation on the stated observation that additional chest deflection measures do not add information to the model), or remove the discussion on the stepwise regression entirely since the deflection measurement correlation in the sled tests is already well documented elsewhere (McMurry et al. 2016).

Response: The intent of the stepwise regression of the individual quadrant deflections was to determine how many of those parameters were necessary, and the findings suggested that two parameters were necessary. Given the stated sample size limitations, an attempt was made to reduce the two parameters to one, both by an additional stepwise regression (which resulted in only the peak overall deflection term), and PCA (as conducted by UVA). The text of the report was modified to clarify these points.

7. When performing regression analyses, it is pertinent to perform post-estimation diagnostics to evaluate the dataset for outliers that are overly influential in the resulting model. McMurry et al. 2016 performed such an analysis on the sled testing dataset evaluated here, and found that one datapoint (the 40 year old PMHS in the 48 km/h lap- belt+airbag condition) was an overly influential outlier. Upon review of that specimen's medical history, it was observed that that subject had suffered from scleroderma, which may have increased the risk of fracture. Because that specimen was a statistical outlier, and because the observed fragility was attributable to and underlying abnormality (scleroderma), that specimen was excluded from that analysis. In the current analysis, was any attempt made to evaluate the potential dataset for outliers, or to scrutinize any potential outliers for abnormalities or pathologies that may warrant their exclusion from the dataset? If such an analysis has not been performed, it should be performed prior to finalizing the IRF.

Response: Given the relatively small sample size, all available data points were retained in this analysis. Additional discussion was included in the Limitations section.

8. Figure 5.10: Age 35 is outside of the age range of the Agency's final dataset (SledOnly), and thus is outside of the range over which the model may be used without extrapolating beyond its underlying dataset. It is likely that the age effect is, in reality, non-linear where the effect of moving towards the younger end of the spectrum levels out around the time of skeletal maturity. In other words, the tolerance to absolute chest deflection (measured in mm) will not continue to increase ad inifinitum as we move younger and younger, but will eventually level off and likely start to decrease in very young

ages where the skeleton is physically smaller. So, it is important to avoid extrapolating outside of range for which your model was developed - especially on the younger end, because we may miss the plateau in tolerance and end up over predicting the increase in tolerance associated with youth. So, to avoid confusion regarding the known range of applicability of the model, please remove the plots for age 35.

Response: The plots were updated to show the risk function evaluated at 40 and 61 years old.

9. Section 5.12 Limitations, the text states "For example, the PMHS are often autopsied after the research test, and rib fractures are investigated at a level of detail that is not possible through physical external examination and reading of radiology from live subjects. Because of this, the number of rib fractures that are recorded from PMHS research tests may be an overestimate of the number of rib fractures that a live human would sustain in a similar loading condition.":

There is a disconnect in reasoning between the first and second sentence above. The first sentence implies that there is a difference in diagnosis sensitivity between clinical observations on living people and rib fractures observed at autopsy in PMHS, which is likely true (the difference between diagnosing via clinical radiology vs. autopsy). The second sentence implies that there is a difference in the actual occurrence of rib fractures between living humans and PMHS. A difference in diagnostic sensitivity does not equal a difference in occurrence. A more accurate statement would be "Because of this, the number of rib fracture that are recorded from PMHS research tests may be an overestimate of the number of rib fractures that would be clinically diagnosed for a living human." Please revise accordingly.

Response: Thank you for the suggested rewording, this was indeed the intent of the text.

10. Figure 5.14: Beyond approximately 40 mm of chest compression, this plot show a greater risk for 7+ fractured ribs than 6+ fractured ribs. This is nonsensical and unrealistic, because 7+ is a subset of 6+. The Agency indicates that this is likely the result of flipping a single data point from injury to non-injury. We suggest either removing the NFR >= 7 IRF due to the fact that it is unrealistic, or evaluating the noted datapoint to determine if it is an overly influential outlier.

Response: The injury criteria does not predict the *number* of fractured ribs, as doing so would require a substantially higher number of observations. Instead, it predicts a probability of a binary outcome, either injury or non-injury. The referenced figure compares discrete risk functions calculated using a different injury definition; as such, it would not be correct to interpret the plot as showing "a greater risk for 7+ fractured ribs than 6+ fractured ribs". Instead, a more accurate statement would be that "above 40 millimeters of resultant chest deflection, using a risk function developed with an injury definition of 7+ fractured ribs shows a greater injury probability than a risk function developed with an injury definition of 6+ fractured ribs." Therefore, the difference in the 6+ and 7+ risk functions is not at all unrealistic, but is simply an artefact of the available sample. Additional explanation of this point was added to the report.

Upon review of the single data point that was referenced in the draft text as contributing to the change in shape, it was determined that there are actually two points that differ between the 6+ and 7+ risk functions, and neither is highlighted by any influence diagnostics. The text was updated to note these two observations as being different between the two risk functions and not overly-influential.

Moreover, the confidence intervals for the risk functions of different injury definitions are largely overlapping, suggesting that the difference in these risk functions is not significant. As NFR >= 7 is consistent with one possible definition used in past research (Laituri et al., 2005), the corresponding risk function will be retained in the report.

Abdomen

R1: For Section 6.5, state the metric that was used to assess biofidelity.

> *Response*: The metric used was BioRank, which is mentioned but only later in the paragraph and not clearly described. A sentence was added to introduce BioRank and point the reader to the appropriate resources for further details.

R1: For Section 6.7, it is unclear why the first paragraph is included. The paragraph concludes that RH data should not be considered independently. This seems to be an unnecessary point since the proposed injury risk combines the RR and RH data.

> *Response*: This paragraph was rewritten to better describe the objective of the nonparametric analysis.

R1: Limitations section cites all of the appropriate injury risks. However, it concludes with '*These limitations should be taken into consideration in the application of the abdominal injury risk function to THOR-50M.*' This statement implies that there are alternative injury risk methods that can be employed or that there is a 'correction' that could/should be made when using this injury risk function. Rather, I believe the intent of this document is to ultimately provide NHTSA's currently accepted injury criterion 'planned for use' per the Introduction. Therefore, recommend removing the last sentence of this section.

> *Response*: The intent of this sentence was to remind the end user that there are limitations to the developed injury risk function that may be relevant to their intended application. For example, those specifically using the abdomen injury risk function to assess injury in a fixed-back application may have more confidence in the results than those using the abdomen injury risk function to assess injury in a free-back application. That said, this sentiment is already suggested by the first sentence of the paragraph, so the offending sentence was removed as suggested.

R5: As mentioned on other metrics, is AIS≥3 the appropriate injury severity? Given vehicle designs today, a lower severity metric would be desirable.

> *Response*: Selection of an AIS level to target for injury risk reduction involves considerations for both the incidence and the severity of the injury, and largely depends on the application. Here, AIS 3+ was chosen because it allowed the most balanced distribution of injury/non-injury observations. That said, AIS 2+ and 4+ risk functions were added to the text for the benefit of others.

R6: Overall, the Agency has come to a reasonable result in the development of the abdominal injury evaluation tools for THOR. The process has some flaws, however, and these should be addressed so that the methodological problems do not undermine the reasonable end result. The feedback below, if implemented, would strengthen the foundation of the proposed criterion and associated injury risk function, but not change them appreciably. The Agency has appropriately excluded penetration rate from the proposed criterion but may encounter resistance based on the historical use of penetration rate in combination with penetration (in the form of V*C or other similar combined metrics). Thus, the Agency's case for not including penetration rate should be made more strongly.

Below is a list of specific comments:

1. The data in Figure 6.2 with regard to mechanism are unreliable and should be removed. The distribution of injury by anatomical structure is reasonable and reliable, but the mechanisms are purely conjecture informed by pre-conceived and unvalidated notions of the CIREN review teams. Having

performed hundreds of thoracoabdominal loading experiments in controlled laboratory conditions, this reviewer is confident that it is not possible in a retrospective review of crash and medical records to bin abdominal injuries by "acceleration", "shear", "compression", and "compression-rate" mechanisms. In fact, the mechanisms asserted in Figure 6.2 are counter to the proposed criterion (penetration as opposed to penetration+rate). The 2008 Kent paper makes a compelling argument for eliminating consideration of penetration rate in the case of seatbelt loading to the abdomen. That experimentally based and well-justified observation should not be diluted by the unvalidated opinions of CIREN case review teams.

Response: As coded in the context of a CIREN case, an injury mechanism of "rate of compression" implies that the injury is not likely to have occurred though static compression alone. This does not suggest that a higher rate of loading would cause a larger risk of injury. A sentence was added to the paragraph describing Figure 6.2 to clarify this sentiment.

2. The Agency's logic for including the RH tests is flawed. As the Agency states, there are only two non-injured specimens in the RH group. Rather than concluding that this does, indeed, indicate that injury propagation may in fact occur while the penetration is held, the Agency fits a statistical model to the isolated RH tests (which they state should not be done), finds that the risk for a given penetration is lower in RH than in RR tests, and uses that finding to conclude that injury was not, in fact, propagating during the hold portion. The problem with that logic, of course, is that any statistical model of the RH tests is flawed by the lack of non-injury tests. There is a very reasonable and logical biomechanical foundation for excluding the RH tests from the injury dataset. Just because the Agency cannot show it statistically does not mean that the tests can be binned – the data are simply insufficient to show the biomechanical reality in the statistics. The RH tests were not designed or intended to be used for the development of injury criteria; they were structural characterization tests. The appropriate interpretation of the Agency's statistical findings with respect to the RH test is that the Agency cannot show that the RH tests and RR tests are different, NOT that the Agency has shown that they are the same.

a. The Agency's desired characteristics in an IRF development dataset (viz. "the dataset includes a combination of both injured and noninjured observations, there is sufficient overlap of these points in the transition region, and there is not a meaningful difference between the logistic and Survival Weibull risk functions") are satisfied using only the RR tests and the data are better from an experimental design standpoint. In fact, if the Agency were to assess the power in a 1-DOF dataset of 33 tests and a 2-DOF dataset of 45 tests, they would probably find that the dataset of 33 tests is better for univariate IRF development than the dataset of 45 tests). In other words, the additional 12 tests are not worth the headache.

Response: Whether or not the RH observations were intended to be included in an injury risk function, they represent valuable observations in a relatively small data set, and, although unlikely, it is conceivable that a similar loading scenario could occur in a motor vehicle crash environment. The analysis described in the report did not intend to form a causal relationship between loading condition and injury as implied by the reviewer, but instead to simply examine the correlation between loading condition and injury risk. Since this could not be done by developing an injury risk function using the RH observations alone, the two risk functions developed using the RR Only and RR+RH data sets were compared. As there were only small differences between the two sets of observations, all available observations were included. One additional clarification which may assist the reviewer's understanding: the risk function formulated using the RR+RH data set did not include loading condition as a covariate, thus the resulting power of a 1-DOF analysis with 45 observations is surely larger than that of a 1-DOF analysis with 33 observations. Further, none of the RH observations appeared to be overly-influential in the risk function formed using the RR+RH dataset.

To address this comment, additional discussion of the RR Only and RR+RH data sets was added to Section 6.9 as well as the Limitations section. Injury risk functions for both the RR Only and RR+RH data sets are now included in Figure 6.8 and Table 6.2 for the benefit of users with specific applications that may or may not include RH-like loading conditions.

3. While the RH tests should be removed from the dataset used for development of the injury risk function, it may be appropriate to use those tests to assess the model's robustness. In some sense, the RH test condition can be considered an alternative "long- time" loading scenario and as such a reasonable check of robustness. Likewise, the Miller tests cited in Kent et al. 2008 would be a useful robustness check. This reviewer suggests using Kent's RR tests to develop the injury risk function and then using the RH tests isolated, the Miller tests isolated, and a combined RH + Miller dataset to assess the robustness of the function. It will show that the injury criterion is a good discriminator in a completely independent dataset and will strengthen the Agency's position.

Response: An analysis was added to the Limitations section to demonstrate this suggested approach. In short, an injury risk function developed using the RR Only data set provides identical injury predictions when applied to the RH observations, as compared to developing the injury risk function using the RR+RH data set. The only difference between the two was a small difference in average error, with the RR+RH risk function resulting in less average error in all but one validation data set. Additionally, several validation data sets were constructed as suggested, and evaluated using both the RR Only and RR+RH risk functions. Only slight differences were found, and in all but one of the validation data sets, the RR+RH risk function resulted in predictions with lower average error.

4. Similarly, the combined dataset of Hardy, Trosseille, and Foster (described in Kent et al. 2008) is another completely independent dataset that should be used in two ways. First, as further confirmation of the validity of the injury criterion and injury risk function by assessing a dataset that is independent and extends the rate of loading beyond that used by Kent and Miller. Second, it is a human cadaver dataset, so the porcine-to-human transformation is partially validated. Again, this strengthens the Agency's position.

Response: Thank you for the suggestion, this was carried out and added to the Limitations section in the discussion of porcine-to-human relationship.

5. Finally, a combined dataset of RH+Miller+Hardy+Trosseille+Foster should be used to evaluate the robustness of the injury criterion and injury risk function. If the AUROC is large (as this reviewer expects it to be), the Agency will have shown discrimination over a wide and diverse range of seatbelt loading conditions and biological models that are completely independent of the tests used for the original development.

Response: Thank you for the suggestion. The RH+F+M+H+T validation data set was added, and the AUROC was indeed large (0.869).

6. The Agency might consider preempting criticism by addressing the issue of penetration rate explicitly in their discussion. The current document is silent on the issue. The Discussion section of Kent's 2008 paper includes a mechanistic justification for exclusion of a kinematic rate term for real-world seatbelt loading.

Response: Section 6.7 was added to introduce the predictor variable (maximum normalized penetration) and briefly describe why it was selected. Additionally, a paragraph was added to the Limitations section to revisit penetration rate with the final model.

7. The Agency might consider discussing the issue of rectus abdominus tensing. That was considered by Kent and could not be shown to modify the relationship between penetration and injury risk. Again, this strengthens the Agency's position.

Response: Thank you for the suggestion, reference to Kent was added in the Limitations section during the discussion on the relationship between PMHS and live humans.

8. The Agency might consider further addressing the fact that the Kent tests were intended to simulate pediatric abdominal loading. The distribution of injury by organ in the swine was validated to a sample of pediatric cases from CHOP, but the pattern of injury is markedly different than that shown in Figure 6.2. Kent addressed this in a limited sense in his 2008 paper by comparing to Miller. The Agency should consider further justification for extending the Kent data to the adult situation. The robustness checks described above accomplish some of that, but there may be additional steps that could be taken to assess that extension of the porcine data.

Response: Again, this suggestion is appreciated and was incorporated into the limitations section through a comparison of the developed injury risk function and the Miller validation dataset. The injury pattern difference is an interesting point, as the CIREN data shows primarily solid organ injury, whereas the Kent 2008 data and referenced field cases show more hollow organ injury. Discussion of this point was added to the Limitations section.

Knee, Thigh, Hip (KTH)

R4: *Section 7.1.* Suggest editing the text "It is important to note that the mechanisms show are regional mechanisms, not organ-specific mechanisms" to clarify that the parts of the KTH are not organs.

> *Response*: Thank you for pointing this out, some of the text was mistakenly copied from the previous section. The sentence in question and the following sentence have been corrected.

R4: S*ection 7.8.4.3.* How was the pairing between neutral posture and flexed posture hip tolerance tests (i.e., the one hip from each PMHS was tested in a flexed posture and the contralateral hip was tested in a neutral posture) accounted for in survival analysis?

> *Response*: The observations from the same PMHS were treated as independent observations in the survival analysis. The text has been updated to clarify this point.

R5: <u>*Knee*</u> - For knee shear condition, the THOR BioRank score was not acceptable, so additional rationale is needed to explain why using the THOR in its present development is acceptable for estimating injury risk.

> *Response*: The knee shear biofidelity condition assesses the sliding joint at the interface between the distal femur and the proximal tibia at the knee, which allows linear translation perpendicular to the tibia, representing both bending of the proximal tibia and extension of the posterior cruciate ligament (PCL). Due to the marginal BioRank assessment in knee shear and the relatively low field incidence of knee shear-related injuries, an injury criterion for knee shear was not proposed for the THOR-50M. The marginal knee shear biofidelity assessment is not believed to influence the femur or hip injury criteria because the knee shear is not a primary component of the load path. This information has been added to Section 7.9.

R5: In the discussion of applied femur force and measured femur force, it is unclear why match pair testing to the Donnelly et al dataset were not performed, since the discussion focuses on this dataset for PMHS and Hybrid III. I also disagree with the author's conclusion that the best approach is to use the Hybrid III value, since they are structurally similar. Based on the earlier discussion of the biofidelity, the THOR femur compliance is much improved and should affect the measured femur force compared to the Hybrid III.

> *Response*: Even if the matched pair testing was conducted using the THOR-50M in the Donnelly et al. 1987 condition, it is still a simplification of the loading condition seen in a motor vehicle crash. Additionally, the relationship between applied force and measured femur force in pendulum impacts is sensitive to both the impact velocity and the pelvis restraint condition. To use this relationship to predict applied force from the measured femur force in the THOR-50M would require knowledge of the impact velocity, which can't currently be measured during a crash test. Instead, the Rupp et al. (2009a) analysis is preferred, as the test apparatus used is representative of the type of loading seen in knee bolster impacts in frontal crash tests. While the THOR-50M femur compliance biofidelity is indeed improved compared to the Hybrid III, the femur compressive element is between the femur load cell and the acetabulum (see Figure 7.4), while the structure of the knee distal to the femur load cell on the Hybrid III and THOR-50M are nearly identical. Thus, the force transmitted from the knee to the femur load cell should theoretically be the same in the Hybrid III and THOR-50M, while the force transmitted to the acetabulum would differ.

R5: *Hip* – In the discussion relating knee force to pelvic force, the same concern arises from the use of the Hybrid III (77%) value instead of conducting additional THOR-M testing to develop a direct transfer function.

Response: As noted above, the force application from the knee to the femur load cell is not expected to be any different between the Hybrid III and the THOR-50M due to the similar structure of the knee. The force transmitted to the acetabulum, however, would differ due to the femur compressive element of the THOR-50M.

R5: Regarding the inability to dynamically measure hip flexion angle (pg 136), would using the components for the acetabular forces allow an estimate of the angle (e.g. if the load is purely in the y-axis, assume a flexion of 0º, etc).

Response: This is a good suggestion, as theoretically if the femur were a pinned-pinned connection between the knee and the acetabulum, the axial force in the femur and the acetabulum reaction forces could be used to calculate the angle of the femur. However, there are several limitations to this simplification that would make this method unreliable. First, the femur can carry moment, which could generate reaction forces at the acetabulum in absence of axial femur force. Second, the femur of the THOR-50M is not a rigid body due to the compressive element between the femur load cell and the acetabulum, which would require additional acceleration measurements to perform the calculations necessary to solve for the femur angle. Finally, as discussed in section 7.7.6, there are additional load paths possible aside from axial femur loading which contribute to reaction forces at the acetabulum.

Lower Extremity

R1: Matched pair tests have not been conducted to determine the Thor-specific injury risk levels. It is stated that '….the risk functions presented here assume that THOR-50M responds exactly like a human cadaver. This is a reasonable assumption given THOR-50M's good to excellent biofidelity…'. While the THOR leg may have good biofidelity for the tested cases, this remains a significant assumption. A final document identifying the methods for THOR injury assessment should include the matched pair test results prior to being approved for final release.

Response: Thank you for the comment. We agree that this is a known limitation of the work presented, and we have noted this limitation in the report. We will consider future testing to improve the tibia criteria, if necessary.

R5: *Lower Tibia* - Recommend reanalyzing risk function using the survival method as well.

Response: A survival analysis was conducted and the resulting risk function was included in the paper.

R5: *Ankle* - Both dorsiflexion and inversion/eversion sections discuss moments as well as angle; however, the THOR-M does not measure ankle moments. I'm interested to know how you would determine ankle moments.

Response: Ankle moment is calculated using the forces and moments measured at the lower tibia load cell and geometric relationships between the load cell location and center of rotation of the ankle.

R6: Overall, the Agency has examined the regional mechanisms of injury to the lower limb and the associated injury criteria. As noted by NHTSA, there are a number of issues when combining disparate data sets of injury and non-injury, differing boundary conditions, and different anatomical sites of load measurement. While NHTSA has made a number of assumptions to address these issues, we would recommend additional justification and analysis for a number of the injury mechanisms and risk functions.

Below is a list of specific comments:

Leg

1. In the evaluation of knee-thigh-hip (KTH) injury risk function, NHTSA thoroughly investigated several approaches to determine the relationship between the peak applied force at the knee and the force at the femur load cell. Figures 7.6 and 7.24 clearly show the effects of inertia on the load distribution along the KTH complex. Unfortunately, NHTSA has not employed a similar methodology for the leg, foot, and ankle complex. In particular, the influence of differing measurement locations and end conditions on the forces and the resulting injury risk curves was not investigated. In Section 8.6.1, NHTSA states that it is not appropriate to include the work of Roberts et al. (1993) that used footplate loads with those of other studies that used loads at the mid- or proximal-tibia. In contrast with this assessment, however, NHTSA combines data sets employing loads measured at the mid- and proximal-tibia without due consideration of the inertial influences. Since the underlying data sets exhibit injury patterns to the distal foot and ankle, we recommend the use of a criterion based on measurements (direct or equivalent) that are only at or distal to the mid-leg. NHTSA's justification in combining data sets with loads measured at different locations is that the resulting mean fracture levels in the disparate tests are roughly equivalent but this does not adequately account for the underlying discrepancies between the studies. If NHTSA wishes to show equivalency of the conditions, they might consider using a human lower limb finite element model to investigate the extent of the inertial effects.

Response: For the Femur criteria, NHTSA actually did combine the PMHS data for forces and fractures occurring at different locations along the femur. The dataset included all fracture locations – proximal, distal, and patella, though most were distal and patella. Then a fracture risk function was developed (for a femur fracture anywhere) vs. the force applied at the knee. In other words, NHTSA did not use the location of the Fx in the development of the risk curve. The transfer function was needed because force at the knee had to be related to force in the femur, not because of differing measurement locations along the length of the femur.

Similarly, in the tibia, only data measured at the tibia was used. If footplate force was to be combined with tibia force in the same risk function, a transfer function would be needed to relate footplate force to tibia force. Inertial effects due to load cell location along the length of the tibia are expected to be small, given than the mass between the proximal and distal tibia is small.

2. For an axial force based criterion that accounts for inertial and rate effects, the injury risk function should be independent of the boundary conditions if the injuries are compressive in nature and there is no eccentricity or curvature involved. NHTSA's justification that the THOR dummy, and humans in real world crashes, is expected to experience both free and fixed boundary conditions of the leg may be true but is not relevant to the injury criteria determination in this case and does not lend credence to combining the data sets with two different load measurement locations.

Response: We agree that this sentence is potentially misleading and have removed. In addition, to further investigate the effect of boundary condition, we investigated the whether a fixed/free covariate would enter the regression model, using criteria of $p<0.1$. We found that the fixed/free covariate did not meet the significance criteria for entry into the model, lending further support for combining all datasets, regardless of boundary condition.

3. In addition to the inertial issues, consideration should be given to the differences between leg (tibia and fibula) and tibia only forces. In particular, the influence of the fibula should be addressed with a conscious decision to either compensate the tibial load measurements or provide a justification for ignoring the effects. According to Funk et al., (2003)[8], the fibula bears an average of 6% of the axial leg load when the ankle is neutrally oriented. The proportion of axial load borne by the fibula, however can change substantially with eversion or inversion. No discussion of these effects are provided in the document.

Response: In response to this comment, we have added a limitation to this section noting that the small amount of load borne by the fibula is not accounted for in the injury risk function formulation.

4. NHTSA's statement "Because many midfoot and forefoot fractures are also caused by loading to the plantar surface of the foot, this criterion (axial load) will likely also address those injuries as well." This appears to be conjecture and should be supported by references that document the levels of load associated with midfoot and forefoot fractures at these loading rates. While there are a number of studies that apply force to the forefoot in order to generate dorsiflexion of the ankle, UVA is aware of a very limited number of impact studies that examine axial load for midfoot and forefoot injuries and none that have developed an injury risk function based on plantar axial load.

---

[8] Funk JR, Rudd RW, Kerrigan JR, Crandall J (2003), ANALYSIS OF TIBIAL CURVATURE, FIBULAR LOADING, AND THE TIBIA INDEX, IRCOBI, 2003.

Response: We agree that this sentence may be misleading, and it was removed.

5. Yoganandan et al. (2014)[9] included repeated testing in select data sets and the associated statistical analysis incorporating repeated samples. It is unclear if any repeated tests were included in the NHTSA analysis.

Response: Yes, there were repeat tests included in the analysis. In response to this comment, a survival analysis was conducted and the resulting risk function was included in the paper. This allowed treatment of repeat tests as interval-censored data.

6. The magnitude of superimposed Achilles loads and its influence on the injury tolerance was not addressed in the analysis despite the fact that a number of the studies explicitly included Achilles tension as part of the experimental design (Kitigawa et al. 1998[10], Funk et al 2002[11]). Funk et al. (2002) demonstrated that fracture initiated at the distal tibia significantly more often in tests with Achilles tension compared to tests without Achilles tension (p=0.049).

Response: While Funk et al. 2002 indeed showed that Achilles tension was significant in the prediction of fracture location, Achilles tension was not significant in the prediction of fracture force ("All predictor variables were statistically significant at the p<0.05 level, except for Achilles tension"). Also, given that the Achilles tension is not known for a given motor vehicle crash occupant, nor is the relationship between the THOR Achilles tension and that of a given motor vehicle crash occupant, the necessary information to include Achilles tension in the risk function is not available. However, given the previous research and this comment, a paragraph was added to the limitations section in the interest of transparency.

7. Yoganandan et al. (2014)[12] determined that a number of samples were overly influential and therefore were removed from further processing. It is not clear if a similar analysis was performed for the data set used by NHTSA.

Response: This was not done. It was desired to include all data, unless there was a physical or biomechanical reason to exclude.

Tibia Moment

8. NHTSA examines the correlation of RTI with tibia moment and axial load. While this is a reasonable assessment in that examines peak values regardless of the time of occurrence, we would recommend examining the percentage of moment and axial load that comprise the RTI as a function of time may be a more effective assessment of the dominant loading.

Response: The intent of the correlation analysis was to examine the relationship between the RTI injury metric and its components, force and moment, to determine whether the RTI presents additional information to improve injury prediction. As the RTI was shown to be highly correlated with resultant tibia moment, use of RTI in addition to resultant moment may not add additional value.

---

[9] Yoganandan N, Arun MWJ, Pintar FA, Szabo A, Optimized lower leg injury probability curves from post-mortem human subject tests under axial impacts, Traffic Inj Prev. 2014; 15(0 1): S151–S156.

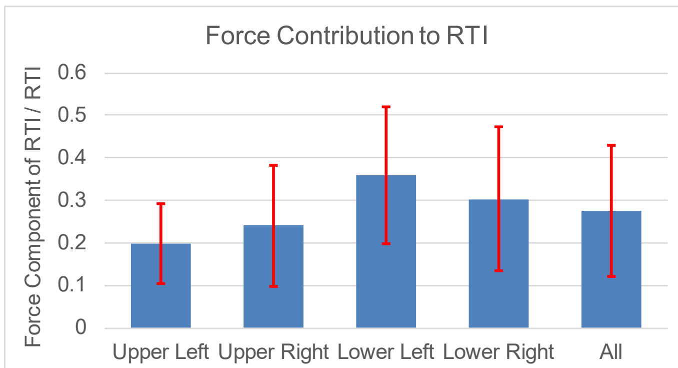
[10] Kitagawa, Y., Ichikawa, H., King, A.I. and Levine, R.S., 1998. A severe ankle and foot injury in frontal crashes and its mechanism. In: Proceedings of the 42nd Stapp Car Crash Conference (No. 983145).

[11] Funk, J.R., Crandall, J.R., Tourret, L.J., MacMahon, C.B., Bass, C.R., Patrie, J.T., Khaewpong, N. and Eppinger, R.H., 2002. The axial injury tolerance of the human foot/ankle complex and the effect of Achilles tension. Journal of Biomechanical Engineering, 124(6), 750-757.

[12] Ibid.

Despite this high correlation, RTI was further analyzed as a potential predictor variable in the updated report.

To the reviewer's suggestion of examining RTI as a function of time, the nature of the calculation and the nonlinearity of lower leg loading in crash tests results in a wide variation in the distribution of moment and axial load, thus using the time-history to examine dominant loading is not straightforward. Instead, the contribution of moment and axial force at the time of peak RTI was calculated for the available fleet tests. The average contribution of the force component of RTI to the total RTI is on average 27%, with lower percentages in the upper tibia and higher percentages in the lower tibia. Only a few observations show a force contribution above 50%, which further suggests that moment is the dominating component to RTI.



Additionally, RTI was calculated two ways: either instantaneously, or using the peak force and peak moment independently. It was found that the two approaches were highly correlated (R^2 = 0.92). Given the similarity, the instantaneous approach is preferred due to the beam theory basis of the RTI formulation.

9. NHTSA contends that while a humanlike induced bending could be calculated based on the applied axial load, variations in the load path due to the centers of pressure at the knee and ankle joints of a human result in potential variation that cannot be adequately accounted for. While variations in center of pressure do exist, it is not clear that ignoring any degree of eccentricity and the associated influence of an eccentric load on bending failure is preferable to the errors associated with variation in joint centers of pressure. A number of studies have examined distribution of pressure centers as a function of joint angle and the eccentricity and curvature is fairly well established. We would recommend that NHTSA perform either an analytical or computational analysis to justify the exclusion of curvature and eccentricity.

Response: This section of the report was completely rewritten to account for the various comments. As noted in response to the question above, RTI is no longer excluded from analysis. In the updated section, RTI was considered as a potential predictor variable.

10. Since the equivalent applied moment due to the externally-applied axial load in the Schreiber et al. (1998)[13] tests was not known, NHTSA used only the tests without superimposed axial load in the formulation of the risk function. It is clear, however, that the axial load does contribute a bending moment due to the eccentricity and the associated moment failure level for Schreiber's tests with axial

[13] Schreiber, P., Crandall, J., Hurwitz, S. and Nusholtz, G.S., 1998. Static and dynamic bending strength of the leg. International Journal of Crashworthiness, 3(3), 295-308.

load is nearly 20% lower than his corresponding tests without axial load. This would seem to provide further support for including some sort of factor related to eccentricity and axial load in the formulation.

Response: We have re-analyzed the Nyquist, Schreiber and Untaroiu datasets to study the influence of axial loading on total moment at failure. In doing so, we have included Schreiber's combined loading data set and used Schreiber's estimate of the axial force contribution to mid-shaft bending of 65 Nm to present a total moment value akin to the total moment from Untaroiu. Survival analysis was run on the datasets from the three sources including all dynamic test data from each. The models evaluated included total moment and Revised Tibia Index. This approach allowed inclusion of all possible data.

Ankle Dorsiflexion

11. The NHTSA CIREN data indicates that ankle malleolar fractures are mostly attributed to compressive axial loading. While malleolar fractures may be caused by axial load, we are unaware of any other studies (biomechanical or epidemiologic) that indicate axial loading as the most common mechanism. In fact, there appear to be numerous studies that indicate rotational measures as the most common injury mechanism. NHTSA's statement on the CIREN data seems to indicate that they support the data showing malleolar fractures are primarily caused by axial load. We believe it would be helpful to provide additional justification for this assertion given that it is not consistent with other biomechanical literature.

Response: The assignment of regional mechanisms in CIREN is informed by biomechanical literature, but is also based on the available physical evidence associated with the injury-producing event. That axial load was identified as the primary mechanism for many malleolar fractures is an indication that the injurious loading conditions in vehicle crashes normally involve axial loading.

Inversion/Eversion

12. The data sets by Petit et al. 1996[14], Parenteau et al. 1998[15], Jaffredo et al. 2000[16], and Funk et al. 2002[17] were designed to impose rotations about a nominal center of rotation. Conversely, the inertial effects of the hardware in Begeman et al. 1993 imposed rotations about an axis not coincident with that for xversion of the ankle. Similarly, the rotation angles to failure appear considerably different between the Begeman data set and those of the other researchers. Any inclusion of the Begeman data set in the evaluation should be justified in light of these differences. UVA recommends removing the Begeman data and reanalyzing the data to see if a more consistent risk function can be developed.

[14] Petit, P., Portier, L., Foret-Bruno, J.Y., Trosseille, X., Parenteau, C.S., Tarriere, C. and Lassau, J., 1997. Quasistatic characterization of the human foot-ankle joints in a simulated tensed state and updated accidentological data. In: Proceedings of the International Research Council on the Biomechanics of Injury Conference, 25, 363-376.

[15] Parenteau, C.S., Viano, D.C. and Petit, P.Y., 1998. Biomechanical properties of human cadaveric ankle-subtalar joints in quasi-static loading. Journal of Biomechanical Engineering, 120(1), 105- 111.

[16] Jaffredo, A., Portier, P., Robin, S., Le Coz, J.Y. and Lassau, J.P., 2000. Cadaver lower limb dynamic response in inversion-eversion. In: Proceedings of the 2000 International IRCOBI Conference on the Biomechanics of Impact, 183-194.

[17] Funk, J.R., Srinivasan, S.C., Crandall, J.R., Khaewpong, N., Eppinger, R.H., Jaffredo, A.S., Potier, P. and Petit, P.Y., 2002. The effects of axial preload and dorsiflexion on the tolerance of the ankle/subtalar joint to dynamic inversion and eversion. Stapp Car Crash Journal, 46, 245- 265.

Response: In response to this comment, Begeman data was excluded from the risk function development, and the justification for exclusion was noted: “Additional inversion/eversion data comes from Begeman et al. (1993), Petit et al. (1996), Jaffredo et al. (2000) and Funk et al. (2002). Of these, the test conditions employed by Parenteau et al., Petit et al., Jaffredo et al., and Funk et al. were designed to impose rotations about a nominal center of rotation. By contrast, the condition used by Begeman et al. applied an off-axis load to the plantar surface of the foot, and did not control for rotation about the center of the ankle joint. Foot plate angles were reported in Begeman et al. as surrogate for ankle rotation, but because of the test setup, these were not equivalent. This is also evidenced by the significantly higher ankle failure angles seen in the Begeman test series, compared with the other studies noted. For these reasons, Begeman data has not been included in further analysis.”

Re-analysis of the risk curves, with Begeman data excluded, is included in the report.

Comparing Fleet and Field Estimated Injury Risk

R4: General comment: NHTSA's argument that estimates associated with small sample sizes in surveys like NASS-CDS are unstable is sound. However, there are methods to address this such as using statistical models to smooth and refine point estimates (e.g., using logistic regression modeling over a larger domain to estimate brain injury risk and then comparing this information to BrIC predictions from crash tests). If the underlying issue is related to the survey design (e.g., excessive weights causing year-to-year variance), there are also ways to handle this, like trimming and redistributing weights, but I don't think this is the issue as nothing in the population of interest should preferentially include cases that commonly have higher weights.

Response: Thank you for your comment. We recognize there are alternative methods to address sample size limitations, point estimates and associated standard errors. After reviewing these methods, we chose a target sample classification scheme from the literature as a benchmark for describing the size of the target population versus the full domain of the sample. The efforts included efforts to maintain a correlated sample requirement over multiple years (15) of NASS-CDS. The sample classification reference was paired a restriction of crash parameters (e.g. delta V, damage extent, belt use, airbag deployment, no rollovers or ejections) that correlated with measured characteristics of full-scale crash tests with the THOR-50M (both NCAP full frontal 35 mph and frontal oblique). Also, as the commenter notes, since we restricted the analysis to severities comparable to the full-scale crash tests, our target population did not include cases with "higher" weights.

General Comments

R4: Appendices - It would be helpful to [add] censoring status in the data tables in Appendices E, F, and G that are used in survival analyses.

Response: Where treated on a case-by-case basis, censoring status was added to the relevant tables in the Appendix (e.g. Appendix G and I). Otherwise, censoring treatment is described in the text; for example, the Survival Analysis section of the Neck chapter (Section 4.9.5) describes that, "All failure data were considered to be left censored and all non-failure data were right-censored."

R5: Please check the document for errors. I ran across a few cases were editorial corrections are needed.

Response: Thank you for the comment. We have gone through and made edits as needed.

R5: The exact validation range of each metric must be documented, so that it is clear if the injury risk prediction is accurate for specific impact conditions / configurations / orientations, etc. Otherwise, they could be misapplied and incorrect conclusions could be reached on injury risk. This is critical to document here in this report, so that the metric isn't misused later.

Response: The exact range of application of each metric is not prescribed, as it depends on many factors specific to the application of the risk functions by the end user. If defined too strictly, the risk functions would apply only to PMHS in laboratory experiments. If defined to broadly, the risk functions could be applied in ways they were not intended. It is left up to the end user to determine whether these risk functions are appropriate, based on the available information, consideration of the documented limitations, and the nature of their application.

R5: Given the proposed use of THOR in oblique tests, are there additional lateral/oblique injury metrics that need to be considered (e.g. neck lateral moment)?

Response: The predominant injury modes seen in small overlap and oblique frontal crashes are head, chest, knee/thigh/hip, and leg injuries[18]. These have been addressed in this report by the inclusion of a brain injury criterion that considers rotational motion including Z-axis rotation; a multipoint thoracic injury criterion that considers not just X-axis deflection but resultant deflection at four possible measurement locations; knee/thigh/hip injury criteria that address both femur fracture and acetabulum fracture in absence of femur fracture, as seen frequently in small overlap and oblique crashes; and leg injury criteria that consider both axial loading and resultant bending moment. It is possible that in the future, additional injury criteria could be developed if sufficient PMHS data became available and field data demonstrate a need to address a given injury mechanism.

[18] Rudd, R., Scarboro, M., Saunders, J., "Injury Analysis of Real-World Small Overlap and Oblique Frontal Crashes," 22nd ESV Conference, Paper No. 11-0384, 2011