CCDC DAC-TR-2020-041
July 2020

W0063 Guidance Document for Human Injury Probability Curves (HIPCs) Development Biomechanics Product Team, Version 1.4

by L. Voo, S. Gayzik, A. Baker, F. C. Hsu, F. Pintar, A. Banerjee, N. Yoganandan, C. Bass, H. Cutcliffe, J. Zhang, J. Rupp, D. Drewry III, M. Montoya, D. Barnes, and K. Loftis

DESTRUCTION NOTICE

Destroy by any method that will prevent disclosure of contents or reconstruction of the document.

DISCLAIMER

The findings in this report are not to be construed as an official Department of the Army position unless so specified by other official documentation.

WARNING

Information and data contained in this document are based on the input available at the time of preparation.

TRADE NAMES

The use of trade names in this report does not constitute an official endorsement or approval of the use of such commercial hardware or software. The report may not be cited for purposes of advertisement.

CCDC DAC-TR-2020-041
July 2020

W0063 Guidance Document for Human Injury Probability Curves (HIPCs) Development Biomechanics Product Team, Version 1.4

by L. Voo, J. Zhang, D. Drewry III, and M. Montoya
Johns Hopkins University Applied Physics Laboratory

S. Gayzik, A. Baker, and F. C. Hsu
Wake Forest University

F. Pintar, A. Banerjee, and N. Yoganandan
Medical College of Wisconsin

C. Bass and H. Cutcliffe
Duke University

J. Rupp
University of Michigan Transportation Research Institute

D. Barnes
SURVICE Engineering Company

K. Loftis
CCDC Data & Analysis Center

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From - To)*
July 2020	Technical Report	2015 – 2019

4. TITLE AND SUBTITLE
W0063 Guidance Document for Human Injury Probability Curves (HIPCs) Development Biomechanics Product Team, Version 1.4

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
L. Voo, S. Gayzik, A. Baker, F. C. Hsu, F. Pintar, A. Banerjee, N. Yoganandan, C. Bass, H. Cutcliffe, J. Zhang, J. Rupp, D. Drewry III, M. Montoya, D. Barnes, and K. Loftis

5d. PROJECT NUMBER
W911QX-17-D-0006

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Director
U.S. Army CCDC Data & Analysis Center
6896 Mauchly Street
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
CCDC DAC-TR-2020-041

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report describes the methods used in generating human injury probability curves (HIPCs) for the Warrior Injury Assessment Manikin (WIAMan) Program. The methods were based on the ISO-18506 and survival analysis with additional statistical tests to include the ranking of the biomechanical injury metrics, three injury-probability functional forms, confidence interval by injury probability level, and the Normalized Confidence Interval Size to determine the curve quality by injury probability level. In addition, an R-studio software code was developed to enable integrated application of the HIPC generation procedure by all WIAMan biomechanics performers.

15. SUBJECT TERMS
Warrior Injury Assessment Manikin, WIAMan, data censoring, injury metrics, injury correlation ranking, injury risk curve, human injury probability curve, HIPC, injury probability curve, survival analysis, Brier Score

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kathryn Loftis
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT	65	19b. TELEPHONE NUMBER *(include area code)* 410-306-0344

Table of Contents

Table of Contents

List of Figures

List of Figures

List of Tables

Executive Summary

This report provides guidance on the methods to be used in the generation of human injury probability curves for the Warrior Injury Assessment Manikin (WIAMan) Program. The document is accompanied by software (SW) program code, written in R, to execute the recommendations contained herein. This SW package contains the tools for each step of this methodology described in this report. The SW can be run using R studio (recommended), which is a graphical user interface and is open source. This package comes with a quick-start guide along with a user manual. The SW package is available to all WIAMan biomechanics performers on the Johns Hopkins University Applied Physics Laboratory-hosted "bioshare" site. Additional details and verification of the methods are provided as appendices.

1. INTRODUCTION

The primary objectives of this working group were to provide 1) technical guidance on the data-collection strategy, 2) analysis methods for injury metrics' ranking and selection, and 3) a working software (SW) code to generate human injury probability curves (HIPCs). The development of the methods for the generation of HIPCs addresses a few key considerations for the state-of-art methodology, curve quality, and Warrior Injury Assessment Manikin (WIAMan) technical requirements. These considerations are defined by a set of framing questions outlined below. The sections that follow address these three framing questions:

1. How do you decide which injury metric is the better predictor?

- Take injury mechanism and severity into account.
- Consider practicality: limit metrics to those that can be translated to the dummy.

2. What is the best statistical solution for the HIPC?

- Prioritize survival analysis due to its advantage in using interval censoring.
- Provide code with flexibility to add other distributions if desired.
- A nonparametric option should be included.
- Use robust distribution checks.

3. What is a measure of quality for a given HIPC?

- It is essential to assess how well the data fit the distribution used.
- It is essential to distinguish between the commonly cited Quality Index and assessment of fit of underlying distribution (ISO, 2014; Petitjean, 2012).
- Include comparison with nonparametric data.

2. SUMMARY OF LITERATURE REVIEW

The relevant literature materials are reviewed and summarized in Table 1. Each study cited is provided with an overview and the general take-away of the study. The work by Petitjean et al. (2009, 2012) and Petitjean and Trosseille (2011), which was later adopted into the 2014 International Organization for Standardization (ISO) TS/18506 document, contains the most comprehensive and state-of-the-art approaches that were adopted as the basis for the methods developed for WIAMan HIPC application.

Table 1. Summary of Biomechanics Literature on HIPC Development

Citation	Framing Question	Brief Overview	Relevancy to HIPC
Kuppa et al. (2003)	1, 3	Provides a methodology on all bullets, used side-impact data from 42 sled tests.	Relevant to automotive HIPC, IARC, and IARV. Addresses decision on dummy metrics.
Wang et al. (2003)	1, 2	Focused on treating the injury sample as interval censored, rather than left. Fits 8 different curve-development methods to various sets of data. Evaluates each under three criteria, one of which is introduced in the study.	Risk curves (RCs) developed treating as interval censored improved RCs developed by max likelihood method. Propose a new criterion based on hazard function to assess best fitting risk curve. They suggest different parametric survival models for different body regions.
Kent & Funk (2004)	1, 2	Used 7 different actual data sets; applied parametric and nonparametric (Consistent Threshold) models. Goodness of fit by adjusted Anderson–Darling Statistic.	Nonparametric model under-estimates risk at low end and over-estimates risk at high end. No parametric distribution was consistently better. Best to design experiments knowing censoring scheme.
Petitjean et al. (2009)	1	Provides a methodology similar to ISO standard, used side-impact tests for WorldSID IARCs.	ISO method: Relevant to automotive IARC and IARV, Survival Analysis with Weibull distribution.
Praxl (2011)	1	Used artificial data to explain different approaches.	Good explanation of statistical processes for biomechanical data. How reliable are IARCs?
Petitjean & Trosseille (2011)	1, 2	Used theoretical data set and defined L2-error. Error was minimized to compare methods.	For uncertain data logistic regression or survival analysis have low errors. The more exact data collected, survival analysis is better. When using low probabilities (5–25%) as thresholds, survival analysis with Weibull is best. Conclusions: "Survival analysis overall leads to the lowest L2 error, regardless of sample size, proportion of exact data, theoretical injury threshold distribution, and stimulus value distribution evaluated."

Table 1. Summary of Biomechanics Literature on HIPC development

Citation	Framing Question	Brief Overview	Relevancy to HIPC
Hasija et al. (2011)	1	Provides a methodology to evaluate risk curves. Looked at well correlated and poorly correlated data sets.	Relevant to analysis of methods, recommendations on methods are provided considering how well data are correlated, how they are censored, etc. Results emphasize the importance of assessing the distribution fit, not just the relative size of the confidence intervals.
Petitjean et al. (2012)	1, 2	ISO/TC22/SC12/WG6 consensus on the definition of guidelines to build injury risk curves.	Applied method, step by step, on building injury risk curves with survival analysis, distribution assessment, and quality checks. Basically, an application of the ISO method for WorldSID 50th male ATD.
Cutcliffe et al. (2012)	1–3	Assesses parametric and nonparametric survival analyses against logistic regression. Introduces Bayesian survival analysis.	Confirmed Petitjean et al. (2012) results with both simulated and real data sets and provides new Bayesian approach to reduce number of overall tests for a "good" risk function, even with relatively poor initial Bayesian priors.
ISO (2014)	1–3	Provides step-by-step procedure to develop risk curves.	A good starting point. Petitjean et al. (2012) applied these methods. The standard released "first edition" 2014 but it has been in development and has been applied well before this.
Yoganandan et al. (2017)	1, 2	Used ISO method for application to foot-ankle tests. Applied "quality index" at specific probability levels.	Used MCW-generated WIAMan data to construct HIPC, IARC, and IARV

Notes: IARC = Injury Assessment Reference Curve; IARV = Injury Risk Reference Value; WorldSID = Worldwide harmonized Side Impact Dummy; ATD = anthropomorphic test device; MCW = Medical College of Wisconsin

2.1 Literature Survey Resolution

While numerous studies were available, there were relatively few that directly addressed the framing questions except for the work by Petitjean et al. A nonparametric option was deemed essential to include in the methods. Survival analysis was chosen as the basis of the outlined methods because it allows the use of interval censoring to allow optimal use of the postmortem human subject (PMHS) specimens. A priority on assessing how well the underlying data fit the distribution was made. The data set by Kuppa et al. (2003) was used as a training and validation set for the development of the methods (Kuppa, 2003). In that study, approximately 42 PMHS sled tests for lateral impact were used to study thoracic injury risk. Later, program data were used from the foot–ankle loading series conducted by MCW.

2.2 Kuppa et al. Data Set

The Kuppa et al. (2003) data set was based on whole-body PMHS side impact sled tests instrumented with chest bands for deflection and accelerometers on various body regions. PMHS were positioned on a sled with a Teflon-coated bench seat and impacted with a stationary load wall, which was padded or rigid and flat, or with an offset at either thoracic or abdominal locations. Impact velocities were low or high at 6.7 and 8.9 m/s. These combinations with repeat tests were done on 42 PMHS—one test per specimen. Injuries were defined as none or Abbreviated Injury Scale (AIS) 3+ for the data set used in the example. Of the 42 samples, 30 had data for all 24 injury metrics to determine the best metric.

2.3 MCW's WIAMan Foot–Ankle Data Set

The PMHS foot–ankle component tests were performed in WIAMan nominal posture (90-90-90) and all with boots. Different velocities were chosen and the same specimen was tested more than once at increasing velocities. Interval censoring was used in those specimens that had both noninjury and injury data points. Censoring was right or left if a specimen only had a noninjury test or injury test. Specimen selection followed stricter WIAMan PMHS inclusion criteria than the Kuppa data set, with only males, controlled bone quality, and close to 50^{th} percentile.

3. DETERMINING THE BEST STATISTICAL SOLUTION FOR THE HIPC

In this section, we review the recommended steps to generate HIPCs and the data required to create them. This methodology is included in the HIPC SW code that accompanies this report.

3.1 Survival Analysis Primer

To understand how HIPCs are generated, it is helpful to begin with an overview of the survival analysis method. Survival analysis is a set of statistical methods for modeling the time it takes for an event of interest to occur. In the field of injury biomechanics, the time variable can be substituted by any monotonically increasing response variable, such as force, moment, or acceleration. For example, an injury risk curve can be generated by modeling the amount of force it takes for fracture to occur. There are other methods to generate injury risk curves (e.g., logistic regression); however, the substantial benefit of survival analysis is that it has the ability to accommodate differing censored observations. In the previous example, the force of fracture for a given specimen can be determined within two bounds from conducting noninjurious and injurious tests. This information yields injury risk curves with additional confidence as compared with other methods. The other advantage is we can ensure the risk curve starts from 0 at time 0 (or no force) using the survival analysis.

In survival analysis, there are three approaches for modeling the injury risk curve: nonparametric (e.g., Kaplan–Meier), semiparametric (e.g., Cox regression), and parametric (e.g., Accelerated Failure Time [AFT]), each with its advantages and disadvantages. The parametric approach was chosen because it functionally relates the biomechanical response variable to injury probability via a fitting procedure to produce an equation that can be evaluated for injury probability at any value of the response variable.

In this fitting procedure, the survival function (which can be directly estimated using the parametric survival model) is a function of the cumulative hazard; that is,

$$S(t) = e^{-\Lambda(t)}, \tag{1}$$

where S(t) is the survival function at time t and $\Lambda(t)$ is the cumulative hazard function at time t. The cumulative hazard function can be interpreted as the sum of the risks one would face going from duration 0 to time t[1]. Any distribution defined for t can serve as a survival distribution. As stated earlier, normally t is the time value in the survival analysis; however, in the field of injury biomechanics, it represents a biomechanically measured variable (e.g., force, moment, strain, and so on). The three commonly used distributions of survival functions are incorporated in the Risk Curve Generator (RCG) code: Weibull, log-normal, and log-logistic. Using Weibull distribution with parameters λ and p as an example, the cumulative hazard becomes $\Lambda(t) = (\lambda t)^p$ and the survival function becomes $S(t) = e^{-(\lambda t)^p}$. Using the lognormal distribution and log-log

distribution, the cumulative hazard becomes the Equations 2 and 3, respectively, although for simplicity the process is elaborated on only using the Weibull distribution:

$$\Lambda(t) = -\log_e\left(1-\Phi\left(\frac{\log_e(t)}{\sigma}\right)\right), \sigma > 0, \tag{2}$$

where Φ is the cumulative distribution function of the normal distribution and

$$\Lambda(t) = \log_e\left(1+(\lambda t)^p\right), p > 0. \tag{3}$$

These survival models can be modified to incorporate a vector of covariates that may affect the injury value. These covariate effects can be studied in this regression form by introduction of a service distribution by considering T = e^Y, so Y = $\log_e T$ (note that log T is replaced for the rest of the report to simplify the notation), where *y* ranges from $-\infty$ to ∞. Treating log T as the outcome variable is more appropriate in this application. If *t* is used as the outcome measure, it is possible to obtain a negative predicted outcome measure, which cannot be interpreted. This would not be an issue if using log T (i.e., Y) as the outcome measure.

With this formulation, a family of survival distributions can be generated by introducing parameter changes of the form below. The covariates can be also incorporated into the AFT model. The AFT model was used as opposed to the Proportional Hazards (PH) model. Although parametric PH models can also be used to relate the covariates to the hazard function, and are applicable for analysis, there are relatively few probability distributions for the survival time that can be used with such models. Additionally, there was the notion of precedent. The approach selected by the HIPC Working Group (WG) was used in the ISO standard based on the work of Petitjean et al. (2011), on which the HIPC methods were based and modified for WIAMan HIPC development. Under AFT models, the covariate effects are measured on the survival time instead of hazard, as we do in the PH model. This feature allows for an easier interpretation of the results because the regression coefficients measure the covariate effect on the mean survival time. The interpretation is similar to that in the linear regression model.

Without loss of generality, one covariate is shown below to demonstrate the Weibull regression model.

$$\log T = Y = \beta_0 + \beta_1 X + \sigma\varepsilon, \tag{4}$$

where β_0 is the intercept, β_1 is the regression coefficient for the corresponding covariate X, ε follows extreme minimum value distribution G(0, σ), and σ is the shape parameter. To link the regression form back to the two parameters—λ and p—for the Weibull distribution, the equation, where λ depends on the covariate, is used. These two parameters—λ and p—can determine the location and shape of the Weibull distribution.

Because this is a parametric model with known distribution, we can directly estimate the parameters (e.g., β_0, β_1, and σ) and their standard errors using iteration algorithms. There is no closed form solution for this problem. These maximum likelihood estimates are calculated using Newton–Raphson algorithm after several iterations.

Furthermore, the predicted value of log T (= $\hat{\beta}_0 + \hat{\beta}_1 X + \hat{\sigma}\varepsilon$) and its 95% confidence interval as [Lower Bound = predicted value of log T – 1.96 * standard error of predicted value of log T, Upper Bound = predicted value of log T + 1.96 * standard error of predicted value of log T] can be calculated. The value of 1.96 is the Z score for the 97.5-percentile point. Here, the standard error of predicted value of log T can be calculated as the square root of the following term:

$$Var(\hat{\beta}_0 + \hat{\beta}_1 X + \hat{\sigma}\varepsilon) = Var(\hat{\beta}_0) + Var(\hat{\beta}_1 X) + Var(\hat{\sigma}\varepsilon) + 2Cov(\hat{\beta}_0, \hat{\beta}_1 X) + 2Cov(\hat{\beta}_0, \hat{\sigma}\varepsilon) + 2Cov(\hat{\beta}_1 X, \hat{\sigma}\varepsilon), (5)$$

where *Var* means variance and *Cov* means covariance. The predicted value and confidence interval of log T can be calculated using *R* statistical software directly. However, because the predicted value for the original T (as oppose to log T) is of interest, an antilog transformation is performed to get the confidence interval in the original scale by simply taking $(e^{LowerBound}, e^{UpperBound})$. As stated previously, this procedure to compute the confidence intervals is performed numerically such that there is no functional form of the confidence intervals.

3.2 Comparison with ISO-18605

The general framework from ISO-18506 has been validated by Petitjean et al. (2011). For the WIAMan program, investigation of the available techniques to assess HIPC risk functions suggests the ISO-18605 framework provides a suitable starting point for ranking the risk functions based on survival analyses.

The ISO/TC22/SC12/WG6 WG of the ISO evaluated the survival analysis methods used by injury biomechanics researchers in the automotive field and suggested a unified and improved approach to generate injury risk curves from retrospective PMHS tests. While the steps in the ISO procedure appear to be detailed, they do not address all nuances of the problem and deficiencies. The HIPC WG identified five specific areas for improvement, which are detailed in Table 2. Furthermore, we summarize in Table 2 the ways in which the WG has addressed these gaps.

Table 1. WIAMan Method Deviations from the ISO Approach

Topic	Gap in Previous Method[a]	WIAMan Method
Choice of biomechanical parameters for the statistical model	Does not provide guidance on what parameters collected in experimentation to use for HIPC	Use Brier Method Score (BMS) to determine parameter that best discriminates injury/noninjury, rank all parameters, chose best statistical fit working from lowest BMS to high.
Threshold for overly influential observations	States use of dfbetas statistic; however, an arbitrary value was used for the cutoff	Use dfbetas statistic based on formula for sample size, $2/\sqrt{n}$
Comparison with nonparametric data	A quantile–quantile (QQ) plot overlay was instructed, but analysis was subjective	Use the Kolmogorov–Smirnov (KS) statistic to determine similarity to nonparametric data and as the basis for selection of the best model fit
Confidence intervals (CIs)	No specific method was provided on how to calculate CIs, although the use of CIs was instructed	Use the methods described in Section 3.1 to calculate CIs
Normalized CI size (NCIS)	These are referred to as the quality index, which the WG feels does not accurately reflect the measure	Name this measure the NCIS, provide capability to report at any required risk level and program CI levels of interest (85%, 90%)

[a] ISO (2014)

First, the ISO approach does not discuss the choice of biomechanical parameters for the model even though it is common to gather more than one metric in any experiment. The Akaike information criterion (AIC) is the suggested method in the ISO approach to choose a statistical distribution. The criterion is an ordinal score that can be used as a simple tool to rank probability distributions such as Weibull, log-normal, and log logistic. AIC scores can, however, overfit in a statistical model and are known in the statistical literature to be asymptotically inconsistent (Findley & Ching-Zong, 2002). AIC also depends on the sample size. Due to missing data, the sample size for each metric can be different. It is not fair to compare AIC across metrics if the sample sizes are not the same. The ISO approach also recommends comparing the chosen distribution with a nonparametric fit through visual examination of the QQ plot and an overlay of the fitted risk curve on a nonparametric risk curve. However, it does not suggest any quantitative measure for comparisons.

Further, the ISO recommends the use of dfbetas statistic to determine overly influential observations/data points. Several versions of this statistic corresponding to different methods of computation are available. The ISO approach suggests a single dfbetas statistic threshold of 0.3 for defining the outlier(s), with no consideration of sample size. The calculation of confidence intervals for survival regression generally relies on asymptotic approximations, for which there are many established methods. From this perspective, the ISO approach does not specify the method to determine confidence intervals. Improvement was made to this procedure for the WIAMan program with further specifications to provide a more robust statistical framework for WIAMan HIPC generation. This effort resolves fundamental ambiguities in the ISO approach

for deriving RCs and demonstrates the feasibility of its use. This procedure can be used in subcomponent and whole-body PMHS tests and for applications in vertical loading environments related to under-body blast events. All statistical analyses described have been implemented in the open-source statistical SW code *R*.

3.3 Input Data

Input data primarily considered in this report come from biomechanical metric(s). These can be an individual parameter such as the peak force, deflection, used by itself or in combination with others. One such combined parameter is the Thoracic Trauma Index (TTI), which uses accelerations from different levels of the ribs and spine, associated with the injury mechanism/experiment under consideration. Other combinations can be considered. The HIPC statistical method developed here is insensitive to particular biomechanical metrics and chooses the best from a statistical point of view. Other viewpoints should also be considered, such as practicality of measurement.

Furthermore, input can potentially include PMHS information such as demographics, external loading scenarios, and injury outcomes. Demographic data such as age and sex, bone mineral density (BMD), body mass index, and certain specimen anthropometry (bone or spinal-column length for subcomponent and stature for whole-body PMHS experiments), can serve as categorical or continuous covariates in future efforts. Inclusion of covariates will be limited by the sample size, and it is anticipated that initial work on including covariates will look at reduced list such as BMD, age, and gender. An analysis of sample-size effect on injury RC is reported in Appendix A.

3.4 Consideration of Data Censoring

Loading scenarios such as acute/single or repeated applications of the insult serve to censor biomechanical metrics. The presence or absence of injury determines the censoring status. Outcomes with no injury are considered as right censored. If the timing of injury is unknown, the outcome is treated as left censored. In contrast, injury outcome is considered as exact or uncensored if the experimental design, instrumentation, and signal analysis allow the determination of the precise timing of injury. If repeated tests are included in the experimental design, interval censoring is used with noninjury and injury outcomes: right censoring associated with the highest severity noninjury test and left censoring associated with the injury test. However, if the repeated testing protocol does not result in injury, data from the highest severity test are used as right censored. The loading and mechanism and severity of injury should remain the same because the underlying risk curve may be different based on the internal mechanics of load transfer to the body region. Treatment of and effects of censoring are discussed further in Section 3.6.2.

3.5 HIPC Statistical Methodology

The following steps (Sections 3.5.1–3.5.6) describe the recommended WIAMan HIPC procedure. The flowchart for this methodology is shown in Figure 1. The words "risk" and "probability" are used synonymously throughout this section.

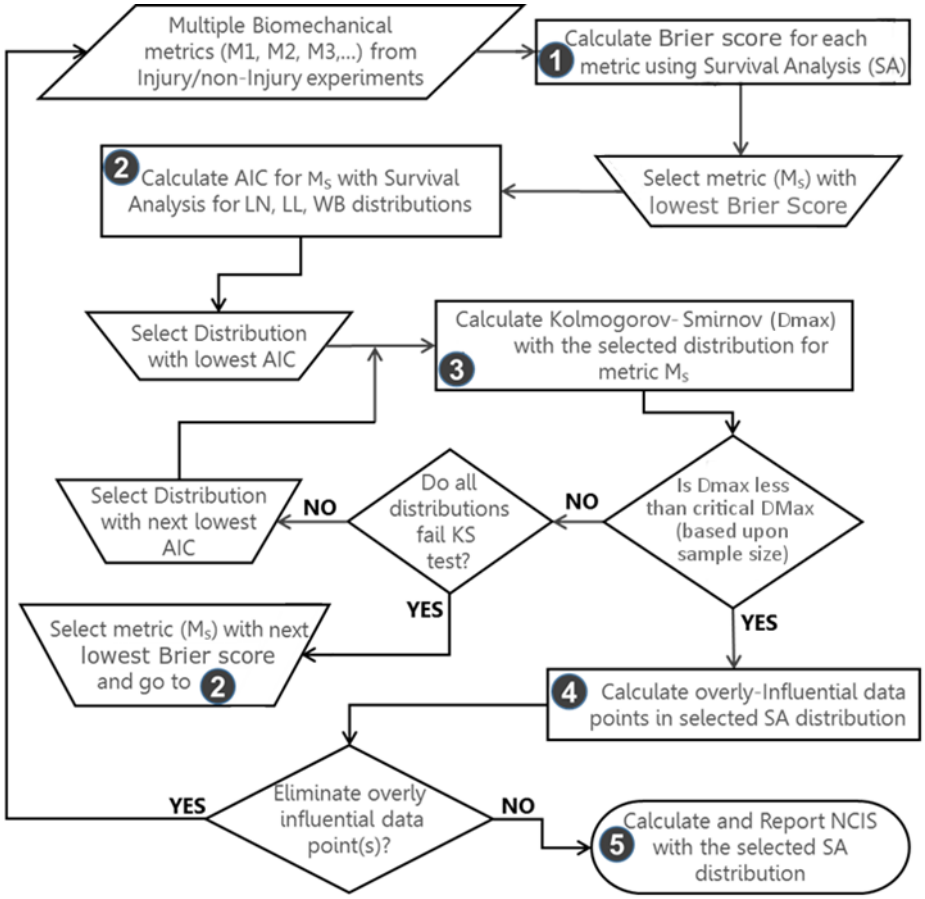


Figure 1. Flowchart of WIAMan HIPC generation process

3.5.1 Choose Optimal Biomechanical Metric Based on Lowest BMS

Though multivariate survival analyses have been developed, survival regression often uses one variable as the time variable. Note that survival analysis is used with time as the time scale in medical statistical analyses, but here the time variable means the biomechanical parameter (e.g., tibia force [Guo & Rodriguez, 1992]). In the presence of several biomechanical metrics, it is therefore important to decide which variable is most likely to produce the best statistical model—the variable best able to distinguish between injury outcomes.

The proposed methodology involves the determination of the BMS for all biomechanical metrics. It represents a measure of the explained variation for survival data at a fixed "time" (i.e., a biomechanical parameter). It can be written as follows:

$$e(\hat{f}_j) = \frac{1}{n}\sum_{i=1}^{n}(I_i(f_j) - \pi(f_j))^2, \quad (6)$$

where $e(\hat{f}_j)$ is the expected error rate, n is the number of human cadavers, $I_i(f_j)$ is the injury outcome of the experiment (either 1 = injured or 0 = uninjured) at or before the value f_j for the biomechanical metric j corresponding to sample i, and $\pi(f_j)$ is the probability of injury at or before value f_j for biomechanical metric j. The $I_i(f_j)$ may or may not be known depending on the censoring status. It will be estimated based on the censoring status (right censored, left censored, and interval censored) and calculated over data informative range (a model independent approach; see Appendix B). The biomechanical metric producing the lowest BMS is the most appropriate metric for further analysis.

3.5.2 Alternative Approach

The area under the receiver operator curve (AROC) is another approach that can be used to select the metric that best discriminates injury risk. It represents a measure of the predictive ability of the metric in terms of specificity and sensitivity. The biomechanical metric producing the highest AROC is the most appropriate metric for further analysis. However, the Brier Method is recommended when interval censored data are included in the data set. AROC also cannot be computed in certain edge cases such as data sets involving only injury samples. Additional details on calculating and interpreting the AROC can be found in Section 3.6.3 and Appendix B. Note that the AROC here simply compares the observed data from the injury probability where injury probability is treated as a risk score.

3.5.3 Choose Optimal Candidate Probability Distribution Based on Lowest AIC

The next step is to choose the most appropriate probability distribution to be used with survival regression, which is typically done under a nonparametric (distribution free) or under a parametric setup. A parametric setup is often advantageous because of its flexibility and generally results in smoother risk models (i.e., accommodation of differing censor status and covariate combinations [Royston & Parmar, 2002]). However, the parametric setup requires additional assumptions: most important, the unknown data-generated distribution is close to the statistical distribution being used. The Weibull, log-normal, and log-logistic are commonly used distributions with survival regression. The proposed methodology calculates the AIC for each of the standard probability distributions used in the parametric survival regression models. The candidate probability distribution with the lowest AIC is identified as the optimal distribution function among the candidates.

3.5.4 Compare AIC Method with KS Tests

The proposed methodology involves calculating the KS D_{max} statistic for that distribution with the lowest AIC. The D_{max} statistic is the basis of the KS test and is the maximum distance between the fitted distribution and the nonparametric fit. Very large values (i.e., $D_{max} > 0.50$)

indicate a poorly fitting distribution; that is, the fitted candidate is significantly different than the nonparametric fit. If the D_{max} value is less than the set criterion, which is usually 0.50 unless there is a very small number of specimens (in which case it is sample-size dependent), the next step would be to identify overly influential observations, described in Section 3.5.5. If D_{max} exceeds the set criterion, the distribution with the next highest AIC is selected (e.g., log-logistic), and the process is repeated. If all distributions fail to meet the criterion, the metric associated with the next lowest BMS is selected and the process is repeated, beginning at Section 3.5.3.

3.5.5 Identify Overly Influential Observations Based on dfbetas Statistic

This involves deleting one observation "at a time" and determining the dfbetas statistic. The statistic corresponds to the difference between the coefficients of the model with the inclusion of all data points and the coefficient with one data point excluded from the ensemble. It can be computed on the basis of any moment or parameter of the chosen probability distribution. The proposed improved methodology involves computing the statistic based on the location, shape, and scale. The excluded data point is considered to be overly influential when the value of the statistic exceeds the magnitude defined by the equation: $2/\sqrt{n}$ where n is the sample size of the ensemble (Belsley, et al., 1980). All overly influential sample points might represent erroneously recorded observations. Therefore, they all should be checked carefully for accuracy by the team that conducted the actual experiments and raw-data analysis. In the absence of errors, the experimental group and the subject-matter expert should decide or confirm the exclusion of any of these overly influential observations, while underscoring that exclusion(s) might lead to leptokurtic distributions and under-represented data variance.

3.5.6 Determine CIs and NCIS

The methodology involves the determination of the CI for the RC obtained in the previous step. The detailed approach for calculating the CI can be found in Section 3.1. The NCIS quantifies the mean, upper, and lower CI curves at discrete probabilities. It is defined as the ratio of the width of the CI to the magnitude of the predictor variable/stimulus from the chosen model at a specific risk level. It is given by Equation 7 (*UL* and *LL* represent the upper and lower limits of the CI and μ represents the mean value at the chosen risk level). A lower NCIS magnitude is associated with a tighter CI at the chosen probability level:

$$NCIS_{risk\,\%} = \frac{(CI_{UL} - CI_{LL})_{risk\,\%}}{\mu_{risk\,\%}}. \quad (7)$$

An R code package has been developed that contains the tools for each step of this methodology. The code can be run using R studio (recommended), which is a graphical user interface (GUI). Details on download and use can be found in the HIPC code package. This package comes with a quick-start guide along with a user's manual.

3.6 Application Consideration of the Recommended Methods

The use of survival analysis techniques is in accordance with the original recommendation due to the inclusion of censored data. Many metrics are routinely gathered and/or derived during postprocessing from PMHS experiments, and the efficacy of the metric describing a complex biomechanical PMHS impact test is generally not known. The first step is, therefore, to choose the most appropriate metric from this pool that predicts injury outcomes. The identification of the optimal metric based on the BMS output is an important and first deviation from the ISO approach. The present statistical analysis assists in the postprocessing phase and may also have implications in designing future experiments of similar complexity and/or loading. More importantly, however, it underscores the need to gather similar data from physical models if the overall experiment is designed to derive injury-assessment risk curves for ATDs, routinely used to improve human safety and vehicle crashworthiness in military environments. The present methodology can, therefore, effectively be used for ATD design, evaluation, injury criteria, and human safety.

3.6.1 Review of Key Definitions and Measures

In the development of this methodology, a number of important definitions and statistical measures were introduced. Each of these is used in Figure 1. In Table 3, we review some of the key statistical measures, their purpose, the critical value, and the origin of the measure in our methodology. Some of these values are used through the process while others are output and, thus, it is important to appreciate the differences. In subsequent sections, we revisit these in more detail.

Table 3. Review of Key Statistical Measures and their Definitions for HIPC Generation

Statistical Measure	What It Does	Critical Value	Origin
BMS[a]	Assess how well a metric predicts observed outcome for censored data	NA, lower the better, ranked by bootstrapping	New method developed in WIAMan program
AIC	Chooses the most appropriate distribution to use for survival regression (Weibull, log-normal, log-logistic)	The probability distribution with the lowest AIC is identified as the optimal distribution function	Commonly reported in parametric survival regression models – adapted to fit the three different test distributions
D_{max} (KS test)	Determines maximum distance between fitted distribution and nonparametric fit	If greater than 0.5, reject	Adapted for WIAMan because we are not using the KS test, rather just the D_{max} value and a critical value[b]
NCIS	Normalized measure of the width of the confidence interval at a specific risk level. (This is an outcome of the code, not a selection factor.)	A lower NCIS magnitude is associated with a tighter CI at the chosen probability level	Adapted from "quality index" to give a more accurate description of the mathematical representation
AROC	Area under receiver-operator curve, similar to BMS but does not work with interval censored data	We do not recommend its use; it is not output by default in the latest version	Commonly used with LR and reported in literature, must be adapted to use w/censored data, replaced by BMS.

[a] Geisser (1993).

[b] For the KS distribution test between two curves, independence between the two distributions is required; but, because the curves are generated using the same data set, they are not independent of each other. D_{max}, the maximum distance between the two curves, is the primary statistic for KS test. It is still calculated meaningfully. However, the variance of D_{max} is not because the dependence between the two curves is not considered. So, the KS test cannot be used as a goodness-of-fit test here. Considering these limitations, the D_{max} statistic itself and critical value of 0.5 were used. This can be interpreted as whether the curves differ by greater than 50% risk prediction at any point.

3.6.2 Censoring Status

The input-data selection involves identifications of likely injury metrics and status, stemming from the same hypothesis, loading paradigm and injury mechanism, and severity. All metrics should be available for each test from the entire experimental data set. The selection of the group of metrics should be decided a priori by the user. In the event of a missing data point for a specific metric, the specimen responsible for the missing metric should not be considered for all metrics or that metric should not be included in the evaluation process. This underscores the importance of gathering the full set of biomechanical metrics for each specimen, a factor not emphasized in the current ISO approach.

The assignment of censoring statuses for injuries and noninjuries is based on sensors used in biomechanical experiments: exact censoring used when injury timing is known from experiments, else, left censoring, and interval censoring for repeated tests on the same specimen.

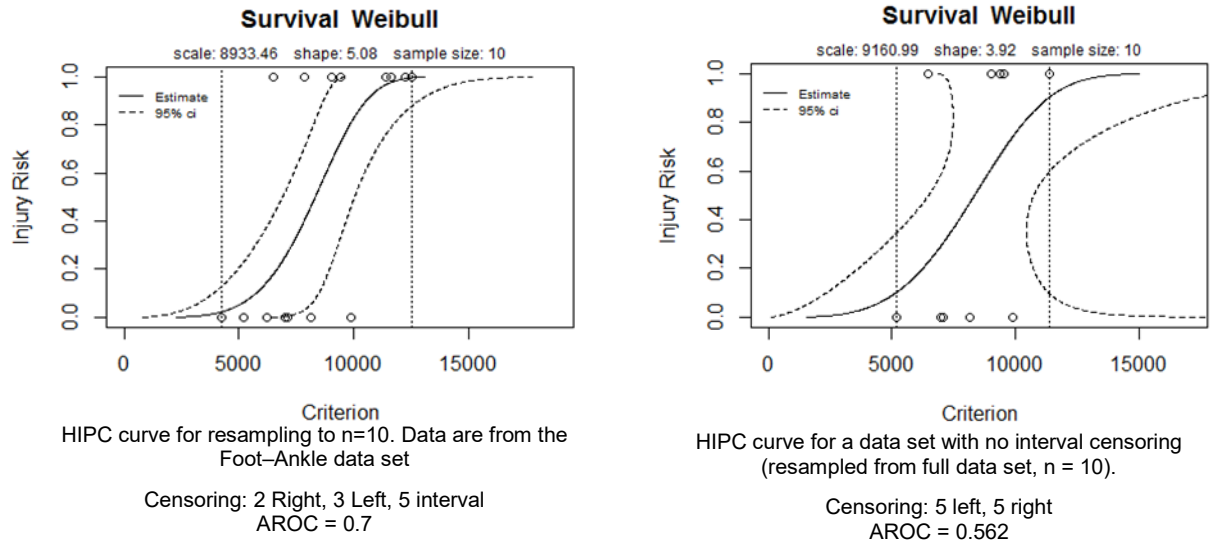
Exact censoring can be used with confidence of instruments such as strain gages and/or acoustic emission sensors are included in the experimental design and the injury is well characterized using these sensors. For example, a strain-gage sensor may provide surface profiles of local deformations of bone; it may not be fully adaptable if fractures initiate from distances farther away from its physical location and to soft-tissue injuries such as ligaments and discs. The biomechanical community has yet to develop a consensus for processing acoustic emission signals, which depend on factors such as rate of load application and type of experimental model (isolated long bone vs. segmented spinal column), because widely varying methods have been suggested based on other disciplines and applied in biomechanical studies (Arun et al., 2014; Cormier et al., 2008; Cormier et al., 2011; Van Toen et al., 2012; Lockner, 1993; McKay & Bir, 2009; Funk et al., 2002; Shridharani, et al., 2014). From this perspective, it will be prudent to resort to the left-censoring scheme to identify injury data points for the development of HIPCs from survival analysis. However, the left-censored assignment for injury data results in lower estimates, a more conservative choice for crashworthiness and safety applications. Censoring status also has significant effects on the AROC and KS, and this is further discussed in Section 3.6.3.

3.6.3 Use of BMS or AROC and KS

The newly introduced procedure using BMS and KS statistics provides important additional statistical information. For example, the BMS is a single measure of the overall performance measure of the model. Other parameters such as the Wald statistic cannot be used for predictive accuracy or goodness of fit (Kuppa et al., 2000). While this statistic has rarely been used in this capacity (Geisser, 1993), extensive internal validation studies are included in Appendix C. With the same underlying sample, it is meaningful to compare BMS between methods and different models, for evaluating performance.

The receiver operator curve (ROC) is obtained by plotting the true positive rate against the false positive rate, with the perfect ROC having an area of unity, while a curve with an area of 0.5 indicates the model is no better at predicting than random guesses. The BMS is recommended rather than the AROC for metric selection although given its historical use, the AROC option is included in the HIPC code. There are a number of limitations with the AROC method. The AROC cannot be computed in certain edge cases such as data sets involving only injury samples. These edge cases are uncommon in previous biomechanical tests. However, if such a situation arises, the adopted experimental design underscores the need to obtain additional data or alter the methodology such that a reasonable combination of injury and no injury data are obtained. If data acquisition precisely points to the time of injury and consist of only injury tests, the AROC cannot be calculated. Since PMHS tests are often designed to extract more than one mechanical metric, as shown in the demonstration data set, other metrics may be considered left censored, which should allow the use of AROC in a carefully planned experimental design.

Interval censoring degrades the AROC and should be considered when choosing the most relevant metrics. Figure 2 shows the effects of having interval censored data. The AROC for the data set is 0.562 (Figure 2 right), and increases to 0.72 (Figure 2 left) when replaced with either left or right censored. This work is expanded upon in Appendix A. Thus, care should be taken when inputting data as interval censored. Additionally, the AROC, due to its formulation, can become undefined when there are no noninjury points (i.e., all data points are left censored). While HIPC curves can still be generated, the AROC cannot, which impedes the selection of a biomechanical parameter. These limitations restrict AROC to ideal situations that may not be present in future experimental tests. The BMS has been shown to have comparable results (Appendix C) under ideal AROC conditions and still performs well under all data sets, including ones that AROC cannot handle. Thus, the BMS is recommended as the metric-selection method in the general case.



HIPC curve for resampling to n=10. Data are from the Foot–Ankle data set

Censoring: 2 Right, 3 Left, 5 interval
AROC = 0.7

HIPC curve for a data set with no interval censoring (resampled from full data set, n = 10).

Censoring: 5 left, 5 right
AROC = 0.562

Figure 2. Effect of censoring on risk curves

The KS measures provide quantitative statistical information (Daniel, 1990). Hence, it was selected to assess the best underlying distribution for the specific biomechanical-injury metric. The KS statistic represents the difference between the fitted distribution and the nonparametric fit, a quantitative measure for the visual overlay plot that is recommended in the ISO approach. These visual plots can be used in the qualitative diagnosis of a model. This feature is also maintained in the present methodology. The chosen critical D_{max} of 0.50 (for sample sizes above 6) represents a reasonable threshold for identifying poor fits, but, the end user can tailor the critical D_{max} to individual requirements. The working group has modified the procedure in the code to not invoke the KS statistical test and associated p-value but rather to use the D_{max} statistic directly as the basis for judging the match between nonparametric and fitted models. The reason for this decision was that the two curves in question are never independent (one is a parametric

fit of the other) and, thus, the test was deemed invalid. After consultation with biostatisticians, there was no alternative test available, and the D_{max} itself was used.

3.6.4 Interpretation of CI vis a vis Generated HIPC

The HIPC CIs provide an assessment of the variance of the mean response. For one input parameter there is a mean percent risk and a confidence about that mean, and based on the HIPC calculation the CI represents the range of input parameters that would result in the same percent output. Based on statistical theory, the true population mean would be within that range for the given level of confidence. The CIs provide a visual representation of the NCIS metric at each point on the mean curve and is an indication of variability in the underlying data. The CIs are included as a representation of the quality of the curve and cannot be used to assess variance of input parameter range at a specific output risk. Large CIs can be the result of large variation or small sample sizes.

3.6.5 NCIS as a Function of Sample Size

Appendix A contains a study on the effect of sample size on the NCIS for various risk and confidence levels. The analysis was run using the data from Kuppa et al. (2003), and each reduction was run with 1000 simulated trials. A summary plot is shown in Figure 3. The conclusion of the investigation was that NCIS levels will likely fall within 1.0 for 85% CI and 1.5 for 90% CI, with sample sizes in the 10-to-15 range at high and low risk levels particularly. While mean NCIS values across all 1000 runs could be considered acceptable, as shown in Figure 3, individual simulated trials do worse. For example, even though for n = 10 at the 5% risk level has a mean NCIS of 0.65, 14.4% of the 1000 simulated runs had an NCIS above 1.0. Additionally, the NCIS varies across different metrics, and the effect of sampling-size reduction is different for other metrics. The effect of metrics, as well as the distribution of NCIS over the 1000 runs, is discussed further in Appendix A.

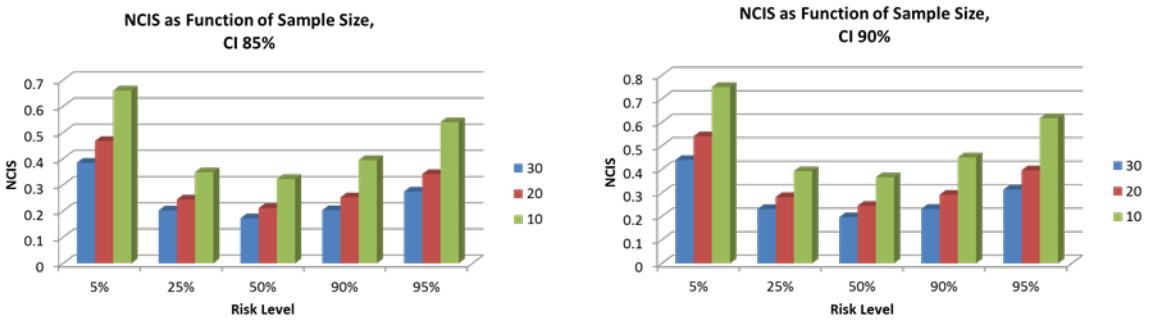


Figure 3. Mean NCIS for different sample sizes and CIs for metric TTI: CI = 90% and CI = 85% both show lower NCIS values as sample size decreases, and overall CI = 85% has lower NCIS values

3.6.6 Interpretation of Overly Influential Observations Identified

The determination and identification of overly influential observations using the dfbetas statistic based on the location, scale, and shape estimates provide additional statistical information and ameliorates its lack of specificity in the ISO approach. In case of excluding data point(s) based on this dfbetas analysis, while the proposed methodology calls for repeating the entire process with the reduced sample size, as a quick alternative, from pure statistical perspective, sensitivity analysis can be done by excluding the overly influential point(s) and comparing the resulting RCs with the curve obtained from using the whole sample.

3.6.7 Specification on CI Calculation

The specification of the approximation method to determine the CI of the injury RC further provides clarity and statistical robustness to the description of injury RCs. The recommended method, outlined in detail in Section 3.1, has been shown to be equivalent to the delta method (Parr, 1983). This process tends to produce wider intervals at the tails, resulting in wider NCIS magnitudes. The risk curve confidence is therefore more conservative than for a conventional logistic regression. This method has been applied in recent studies as an extension of the ISO recommendation (Yoganandan et al., 2016, 2017).

3.6.8 Quality Index vs. NCIS

Petitjean et al. use the term "quality index" in their description of the injury RC methods. ISO simply includes a Quality Index based on relative size of the 95th CI. This is problematic. Quality should include an assessment of the distribution and some assessment of the relative size of the CI at desired points. A methodology is provided here that addresses the goodness of fit of the RC, checks it against a nonparametric curve, and separately evaluates what was formerly referred to as the "quality index."

This report purposely chose to use the term "normalized confidence interval size" rather than "quality index," which is a more accurate description of the mathematical representation of this index. NCIS is merely a normalized measure of the width of the CI at a specific risk level. The NCIS will vary at risk levels based on the data used in the construction of the RC. The HIPC code provided output NCIS values at the relevant risk levels for CIs of note to the program.

3.6.9 Confidence in Selected Metric as Metric Rating

Bootstrapping has been employed as an option in the metric-selection code to evaluate the likelihood that the selected metric is the most discriminating metric among those in the data set. The default value in the code is bootstrapping = false, so unless the user specifies, bootstrapping is not performed. If the user specifies bootstrapping = true, the HIPC metric-selection methodology is performed on each bootstrapped data set and the best metric is selected. This

process is repeated 1000 times by default, but can be set to a user-defined value; the percentage of times each metric is selected as the most discriminating metric is recorded. This percentage is referred to as the "Metric Rating."

It is recommended that relative metric ratings be compared. If several metrics have similar Metric Ratings, each of these metrics should be considered equally viable. If the sample size for the selected metric is less than 10, the bootstrapping approach may not be reliable. For more information, consult the HIPC Risk Curve Generator code's quick-start guide.

3.7 Consideration of Covariates

Often, models can be adjusted using additional information about each observation (covariates). Including covariates in the HIPC generation will slightly change the HIPC methodology. Most notably, covariates can be included in the initial metric selection of the model. While distributional models can be generated, the actual curve cannot be generated unless a particular covariate value is assumed (e.g., discrete curves at ages of 30 and 40 years or for male and female).

Because the curve cannot be generated unless a particular covariate value is assumed, the KS cannot be calculated. The HIPC code package (v. 3.1.4) has preliminary support for covariates, which included criteria for inclusion or exclusion of covariates based on the likelihood ratio test and Wald test. The covariate effect on the injury status will be calculated. Once validated, such criteria will be included in the methodology for HIPC and IARC generation. Two such criteria that have been identified so far are the following:

1. *Likelihood ratio test*—valid in identifying which covariates to include in the model by comparing the model with covariates and the model without covariates. If the test is significant ($p < 0.05$), it means the covariates are associated with injury risk and should be included in the model. We support the use of this statistic in assessing the significance of covariates.

2. *Wald statistic*—also commonly used by statisticians. However, at the sample sizes expected within the WIAMAN program, the likelihood ratio test performs better and is more stable at smaller sample sizes. Thus, the likelihood ratio test is the recommended statistic in determining covariate significance.

These two criteria address use of the likelihood ratio test, as determined by the chi-squared distribution with the appropriate degrees of freedom, and the use of the Wald statistic.

3.8 Methods for Injury Metrics Combination

3.8.1 Motivation

More complex injury outcomes in the WIAMan HIPC's development call for multiple biomechanical-injury metrics to account for their multiple injury mechanisms. Statistical analysis of the data has shown that more than one injury metric could have high correlation to injury status while their linear combination could improve such correlation. Sections 3.8.2–3.8.4 describe the process of determining which metrics to combine and how to combine them.

3.8.2 Recommended Methodology Overview

A multistep approach is recommended if multiple metrics may be combined to create an improved metric. Metric combination should be performed before the metric-selection process and the resulting combined metric(s) should be used as metrics fed into the standard HIPC metric-selection process using the HIPC code. This process is reflected in the flowchart in Figure 4. The combined metrics module is included in the latest version of the HIPC code, Version 3.1.4. For the subsequent recommendations, data were used from the Kuppa et al. data set because of the large number of metrics reported and a relatively large number of observations. The benefits of the combined metric approach are summarized in Table 4.

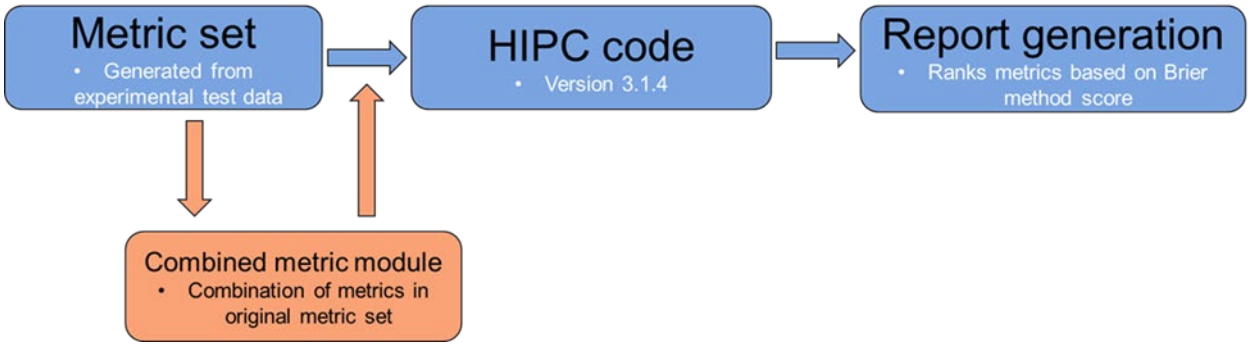


Figure 1. Workflow with combined metric module (orange); original HIPC workflow remains in blue

Table 2. Summary of Metric Combination Methodology and Associated Benefits, Broken Down by Gaps and Proposed Solutions

Existing Gap	Proposed Solution
Current code uses only single metric	Add module for metric combination
Candidate metrics for combination may be at different scales	Normalize metrics by sample's mean. Standard deviation normalization is an option but mean is the default.
No existing recommendations on determining number of metrics to combine	We have selected two. Not recommended to include more due to overfitting concerns.
No existing recommendations on methodology to combine metrics	Summation of normalized metrics with weighting factors: • Weighting factors determined by principal component analysis (PCA) • Weighting factors set to unity
No existing recommendations on how to evaluate combined metrics	BMS
No existing recommendations on which combinations should be evaluated	All combinations should be considered due to computational cost being low. Utilize the BMS as a filter for including into the normal metric-selection process. Provide recommendations for biomechanists to use suggested combinations that have improved BMS.

3.8.3 Metrics Combination Methods

Because many of the metrics operate at different scales (i.e., some are orders of magnitudes larger than others), normalization of each metric should be performed before combination. A linear normalization should be applied by dividing by the sample's mean. Standard deviation normalization is also included as an option in the code but normalizing by the mean value is the default.

Evaluation criteria for combined metrics are recommended as the BMS, and if a combined metric has a lower BMS than an existing metric, that metric should be considered along with existing metrics. Assuming all possible combinations of two metrics are to be considered in the code, there should be $\binom{n}{2} = \frac{n!}{2(n-2)!}$ combinations, where n is the number of metrics to be combined and when considering only pairwise combinations (two metrics):

$$CM_n = \sum_i w_i \cdot \frac{M_i}{\mu_i} \; or \; \sum_i w_i \cdot \frac{M_i}{SD_i} \tag{8}$$

and

$$CM_n = \sum_i \frac{M_i}{\mu_i} or \sum_i \frac{M_i}{SD_i} \tag{9}$$

where CM_n is the n^{th} combined metric, w is the weight, M is the metric, μ is the mean, and SD is its standard deviation. The subscript i in our methodology has a maximum value of 2 because we have capped combined metrics at 2 due to overfitting concerns, although more metrics could be combined. The weights, w, in Equation 8 result from PCA. PCA uses an orthogonal

transformation to summarize intercorrelated variables into a smaller set of components according to their underlying relationships among variables. The resulting components are uncorrelated. Each component is a linear combination of the original variables. The transformation is defined by a set of vectors of weights or coefficients that map each original variable to a new principal component. Equation 9 shows a special case where weights are assumed to be unity and PCA is not used. Normalization using the mean or standard deviation controls the variance in each metric and also makes the metrics unit less. This is important because when adding metrics, a metric with a much larger variance would be overly represented in the combined metric. Normalizing by standard deviation has precedence in normalization for PCA, along with the calculation of z-scores (Härdle & Simar, 2007). However, in the field of biomechanics, mean normalization of metrics is a common method and, thus, our approach is to use mean normalization unlike the conventional z-score methods (Knopman et al., 2016). Additionally, an increment of the normalized metric can also be interpreted as corresponding to a percentage of the data. For example, the normalized mean ± 2, given a normally distributed metric, would represent 95% of the data.

The code available in R incorporates this combination process. In the HIPC code (version 3.1.0 and later) there is a separate module to try up to two combined metrics. Additionally, the combined metrics are written out as HIPC input files such that the user can incorporate them into normal metric-selection process.

The weighting factors can be determined in one of two ways. The first is through PCA, by taking the largest principal component and using the associated weights from the largest component. The second is to simply assume the weights are unity. Given the relative lack of research in this area, we thought it was prudent to code in both approaches, but unity weighting is the default method in the code (users can define which method they would like to choose).

This may allow false positives; there will be testing of many different combinations and it may be due to chance that the combined metric performs better than the original metrics (singular metrics). However, there are no existing statistical tests for significance testing of the BMS (i.e., is the BMS better by incorporating additional metrics, or is the BMS simply better by chance), and it is left to the HIPC code user to judge whether combined metric represents a potentially physically significant quantity. Furthermore, should two metrics appear to be beneficial to combine, the code user can use that information to develop ad hoc combinations through their own regression analyses, which can then be tested in the main code.

3.8.4 Application Examples

Worked examples of metrics combination using the HIPC code are presented here using the Kuppa et al. data. In this data set, there are 30 left- and right-censored observations and 25 different metrics. Some metrics, such as TTI, are already combined metrics. For the mean

normalization results, using either combination method (PCA or summing) resulted in the same top 10 metrics selected, and TTI is included in the best four combined metrics (Figure 5).

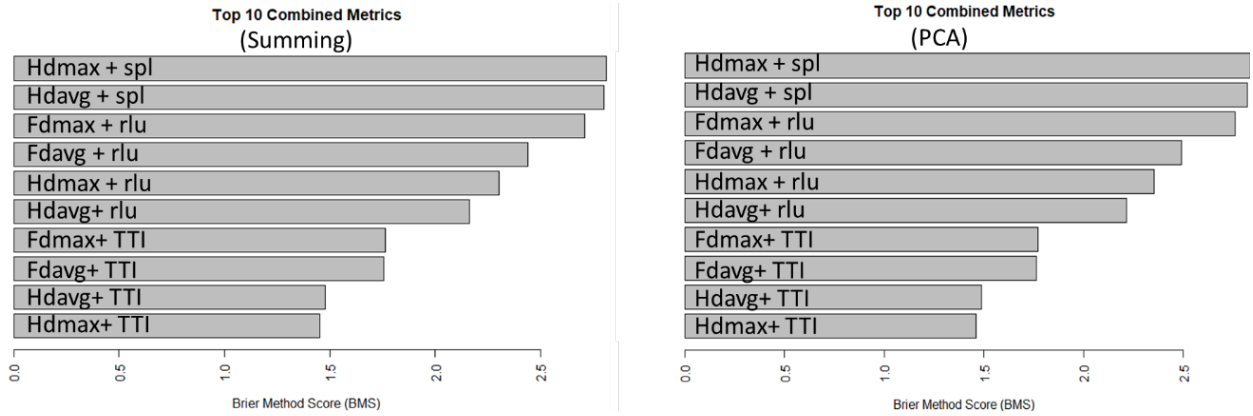


Figure 2. Mean normalization results showing top 10 combined metrics (lower BMS is better)

For the standard-deviation results, a different set of 10 metrics was selected depending on which combination method (PCA or summing) was used. TTI is included in two out of the top four combined metrics (Figure 6).

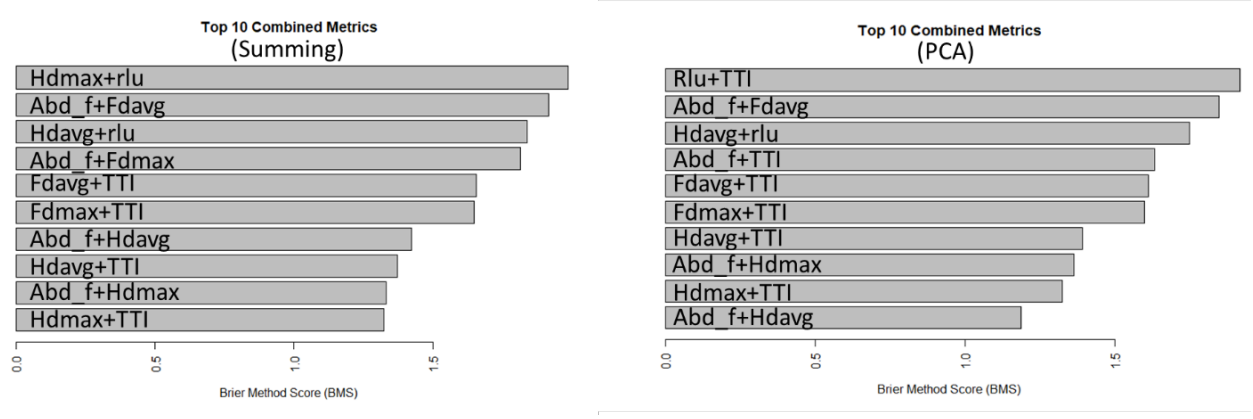


Figure 3. Standard-deviation normalization results showing top 10 combined metrics (lower BMS is better)

This section highlights an example of using the HIPC code to explore the metric combination approach. When normalizing by the mean, both summing and PCA resulted in the same top 10 metrics. Using standard-deviation normalization, the results are slightly different between the two methods. From these metrics, the user can explore using the combined metrics in the metric-selection code if desired. At the time of developing this technique, available WIAMan project data were limited; therefore, no definitive answer was reached as to which method of normalization was superior. Both methods are included in the HIPC code but the default method is mean normalization.

3.8.5 Summary

The users are recommended to use the approaches outlined in this report to identify potential metrics to be combined or investigated for further analysis. This process is included in the HIPC code, which outputs a summary report along with various combined metrics input files. The two methods (sample mean vs. sample standard-deviation normalization) diverge only in how the weights for summation are determined; further, the analysis shows good overlap between the results, with four of the five best combined metrics by BMS being the same in each approach. The user will be able to toggle between the two methods. While the combined metrics can be directly utilized in the metric-selection process, the user should consider the selected combined metrics identified in the summary report as a starting point for developing more complex combinations of metrics, perhaps through regression equations. The methods in the HIPC code consider all combinations but add negligible computational time. The users are, however, advised to decide from all possible combinations and downselect those biomechanically more useful or usable for the development of IARC for a particular ATD where the necessary measurements and calculations can be accomplished.

4. DISCUSSION AND CONCLUSION

This technical guidance document describes the state of the art of the injury risk-curve generation methods, the adoption of the best existing methods and creation of new methods, and validation of the recommended methods for WIAMan HIPC development. Most of the testing and validation of the methods were performed using non-WIAMan biomechanics data either from the open-publication or unpublished data from the participating research laboratories, as the WIAMan HIPC data were being generated and analyzed while the methods were being created and documented in parallel to meet the WIAMan's product-development schedule. Nevertheless, the core methods for the HIPC developmental procedure have been rigorously tested and heavily used in the generation of all of the delivered WIAMan HIPCs for foot–ankle, tibia, femur, pelvis, lumbar, and head–neck anatomies. In addition, the generation of the WIAMan IARC employed the same core procedure and statistical tests as well as the HIPC code. The successful development of the many HIPCs and IARCs for WIAMan is a testament to the impact this report has had and will make to the protection of our Warriors in the years to come.

5. REFERENCES AND DOCUMENTS

Arun, M. W., Yoganandan, N., Stemper, B. D., & Frank A Pintar. (2014). A methodology to condition distorted acoustic emission signals to identify fracture timing from human cadaver spine impact tests. *Journal of the Mechanical Behavior of Biomedical Materials*, 40, 156–160.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.

Bi, J. & Bennett, K. (2003). *Regression error characteristic curves* [paper presentation]. Proceedings of the Twentieth International Conference on International Conference on Machine Learning (ICML 20003). AAAI Press, 43–50.

Cormier, J., Manoogian, S., Bisplinghoff, J., McNally. C., & Duma, S. (2008). The use of acoustic emission in facial fracture detection. *Biomedical Sciences Instrumentation*, 44, 147–152.

Cormier, J., Manoogian, S., Bisplinghoff J., Rowson, S., Santago, A., McNally, C., Duma S., & Bolte, J., 4th. (2011). The tolerance of the maxilla to blunt impact. *Journal of Biomechanical Engineering*, 133(6), 064501.

Cutcliffe, H. C., Schmidt, A. L., Lucas, J. E., & Bass, C. R. (2012). How few? Bayesian statistics in injury biomechanics. *Stapp Car Crash Journal*, 56, 349–386.

Daniel, W. W. (1990). *Kolmogorov–Smirnov one-sample test*. Applied Nonparametric Statistics, 2nd ed., 319–330. PWS-Kent.

Findley, D.F. & Ching-Zong, W. (2002). AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autotregressions and related models. *Journal of Multivariate Analysis*, 83(2), 415–450.

Funk, J. R., Crandall, J. R., Tourret, L. J., MacMahon C. B., Bass, C. R., Patrie J. T., Khaewpong, N., & 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. https://doi:10.1115/1.1514675

Geisser, S. (1993). *Predictive inference: An introduction*. Chapman and Hall.

Green, D. M. & Swets J. A. (1966). *Signal detection theory and psychophysics*. Wiley and Sons.

Guo, G. & Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *Journal of the American Statistical Association*, 87(420), 969–976.

Härdle, W. K., & Simar, L. (2007). *Applied multivariate statistical analysis* (Vol. 22007). Springer. https://doi.org/10.1007/978-3-540-72244-1

Hasija, V., Takhounts, E., & Ridella, S. (2011). *Evaluation of statistical methods for generating IRC*. 22nd International Technical Conference on the Enhanced Safety of Vehicles (ESV), Washington, DC.

[ISO] International Organization for Standardization. (2014). *Procedure to construct injury risk curves for the evaluation of road user protection in crash tests (ISO/TS/18506)*.

Kent, R. W. & Funk, J. R. (2004). *Data censoring and parametric distribution assignment in the development of injury risk functions from biomechanical data (SAE Paper No. 2004-01-0317)*. Society of Automotive Engineers (SAE) World Congress, Detroit, MI.

Knopman, D. S., Gottesman, R. F., Sharrett, A. R., Wruck, L. M., Windham, B. G., Coker, L., Schneider, A. L. C., Hengrui, S., Alonso, A., Coresh, J., Albert, M. S., & Mosley, T. H., Jr. (2016). Mild cognitive impairment and dementia prevalence: the Atherosclerosis Risk in Communities Neurocognitive Study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2, 1–11. https://doi.org/10.1016/j.dadm.2015.12.002

Kuppa, S., Eppinger, R. H., Maltese, M., Naik, R., Pintar, F. A., Yoganandan, N., Saul, R., & McFadden, J. (2000). *Assessment of thoracic injury criteria for side impact*. Proceedings of the International Research Council on Biomechanics of Injury; Montpellier, France.

Kuppa, S., Eppinger R. H., McKoy F, Nguyen T., Pintar F. A., & Yoganandan N. (2003). Development of side impact thoracic injury criteria and their application to the modified ES-2 dummy with rib extensions (ES-2re). *Stapp Car Crash Journal*, 47, p. 189–210.

Lockner, D. (1993). The role of acoustic emission in the study of rock fracture. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, 30(7), 883–899.

McKay, B. J. & Bir C. A. (2009). Lower extremity injury criteria for evaluating military vehicle occupant injury in underbelly blast events. *Stapp Car Crash Journal*, 53, 229–249.

Parr, W. C. (1983). A note on the jackknife, the bootstrap and the delta method estimators of bias and variance. *Biometrika*, 70(3), 719–722.

Petitjean, A., Trosseille, X., Petit, P., Irwin, A., Hassan, J., & Praxl, N. (2009). Injury risk curves for the WorldSID 50th male dummy. *Stapp Car Crash Journal*, 53, 443–476.

Petitjean, A. & Trosseille X. (2011). Statistical simulations to evaluate the methods of the construction of injury risk curves. *Stapp Car Crash Journal*, 55, 411–440.

Petitjean, A., Trosseille, X., Petit, P., Irwin, A., Hassan, J., & Praxl, N. (2012). Injury risk curves for the WorldSID 50th male dummy. *Stapp Car Crash Journal*, 56, 323–347.

Picard, R. R. & Cook R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387), 575–583.

Praxl, N. (2011). *How reliable are injury risk curves?* 22nd International Technical Conference on the Enhanced Safety of Vehicles (ESV), Washington, DC.

Royston, P. & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15), 2175–2197. https://doi.org/10.1002/sim.1203

Shridharani, J., Schmidt, A., Cox, C. A., Bigler, B., & Knight, A.E., & Bass, C.R. (2014). *Dynamic failure localization in spinal specimens using acoustic emissions* [paper presentation]. 2014 IRCOBI Conference Proceedings–International Research Council on the Biomechanics of Injury, 166–175.

Till, D. & Hand R. (2012). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45, 171–186.

Van Toen, C., Street, J., Oxland, T. R., & Cripton, P. A. (2012). Acoustic emission signals can discriminate between compressive bone fractures and tensile ligament injuries in the spine during dynamic loading. *Journal of Biomechanics*, 45(9), 1643–1649.

Wang, L., Banglmaier, R., & Prasad, P. (2003). *Injury risk assessment of several crash data sets (SAE Paper No. 2003-01-1214)*. SAE 2003 World Congress & Exhibition, Detroit, MI. https://doi.org/10.4271/2003-01-1214

Yoganandan, N., Banerjee, A., Hsu, F. C., Bass, C. R., Voo, L., Pintar, F. A., Gayzik, F. S. (2016). Deriving injury risk curves using survival analysis from biomechanical experiments. *Journal of Biomechanics*. 49:3260–7.

Yoganandan, N., Chirvi, S., Voo, L., DeVogel, N., Pintar, F. A., & Banerjee, A. (2017). Foot-ankle complex injury risk curves using calcaneus bone mineral density data. *Journal of the Mechanical Behavior of Biomedical Materials*, 72:246-251.

Zweig M. H. & Campbell G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*. 39(4), 561–577.

Appendix A – Effect of Input Data on Risk Curves and Normalized Confidence Interval Size (NCIS)

Various studies have been undertaken and reported on how the data input to the method can change the resulting risk curves and, in particular, the width of the confidence intervals. We have found that the number of data points as well as the censoring status of the data points affect the risk curves.

A.1 Degradation of AROC under Sample Sets with Imbalance of Injury Status and Censoring

The area under the receiver operator curve (AROC) is affected by a mix of injury and noninjury points. If there are not enough injury points or noninjury points, the AROC is not a reliable differentiation among discriminating metrics. Additionally, if the data set is all injury or all noninjury, the AROC is undefined and has no meaning.

Table A-1. Kuppa et al. Data Simulate Not Having Good Mix of Injury–Noninjury Data Points

# (Injury)	# (Noninjury)		Hdmax AROC	Hdavg AROC	TTI AROC
24	5		0.85	0.833	0.8
24	10		0.8625	0.8625	0.8
24	15		0.8861	0.8778	0.7937
24	18	allSubs	0.8611	0.8704	0.8228
15	18		0.9296	0.9259	0.8593
10	18		0.9167	0.9	0.8333
5	18		0.9556	0.9667	0.9222

Censoring status also influences the AROC. Interval censoring degrades the AROC and should be considered when choosing most relevant metrics. Figures A-1 and A-2, featuring human injury probability curves (HIPCs), show the effects of having interval censored data. The AROC for the data set is 0.562, and this increases 0.72 when removed and replaced with either left or right censored. Thus, care should be taken when inputting data as interval censored.

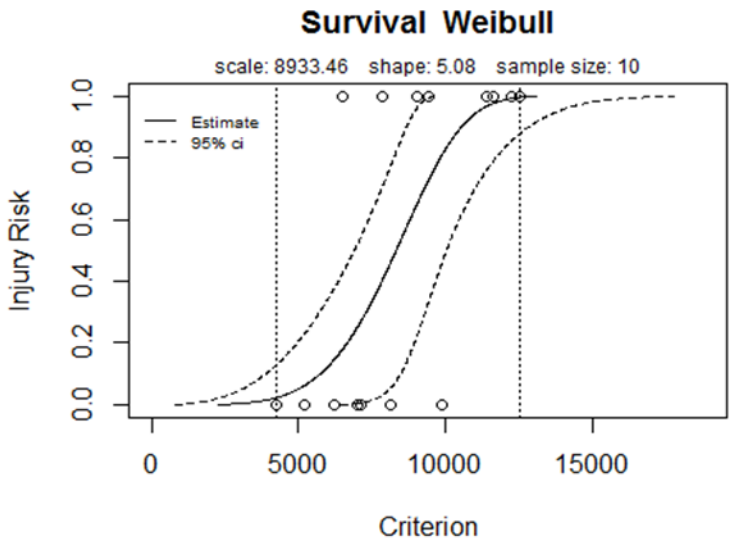


Figure A-1. HIPC for resampling to n = 10; data are from foot–ankle data set (censoring is 2 right, 3 left, 5 interval)

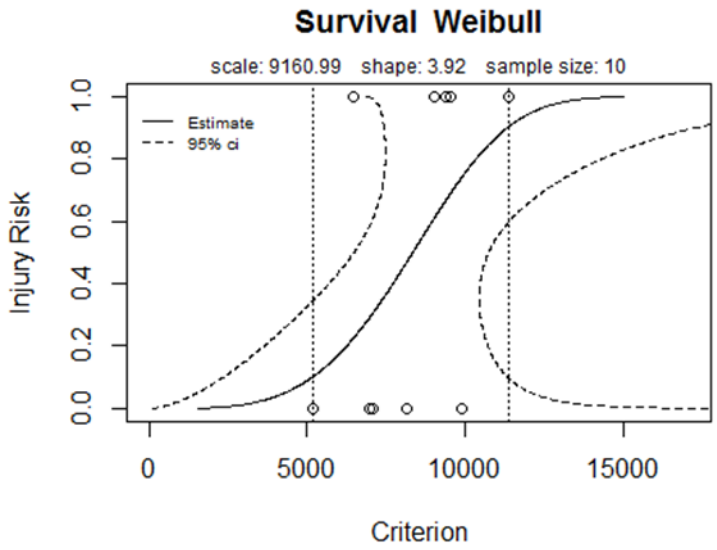


Figure A-2. HIPC for data set with no interval censoring (resampled from full data set); data are from food–ankle data set, censoring is 5 left and 5 right

A.2. Quality Parameters' Degeneration with Smaller Data Sets and Worse Metrics

Quality measures degrade with smaller sample sizes and across different metrics. An overview of the quality degradation is shown in Figure A-3. Figure A-4 shows the Thoracic Trauma Index (TTI) down-sampled from 30 samples to 20 samples, and Figure A-5 shows 10 samples. TTI was considered the best metric in this data set based on AROC.

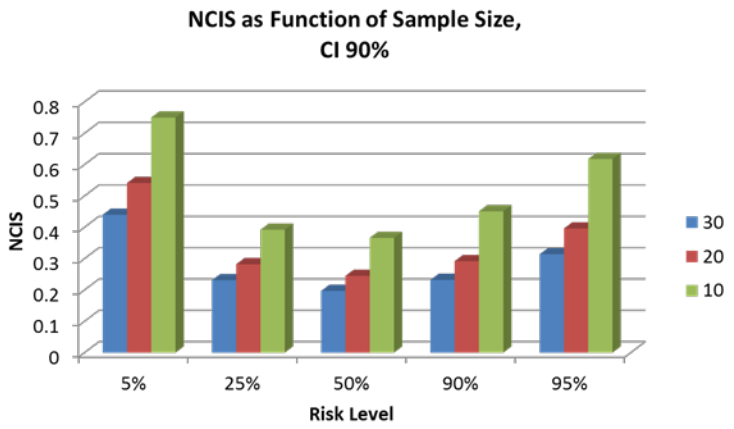


Figure A-3. Mean NCIS for parameter TTI (1000 simulated trials) for different sample sizes at CI of 90%

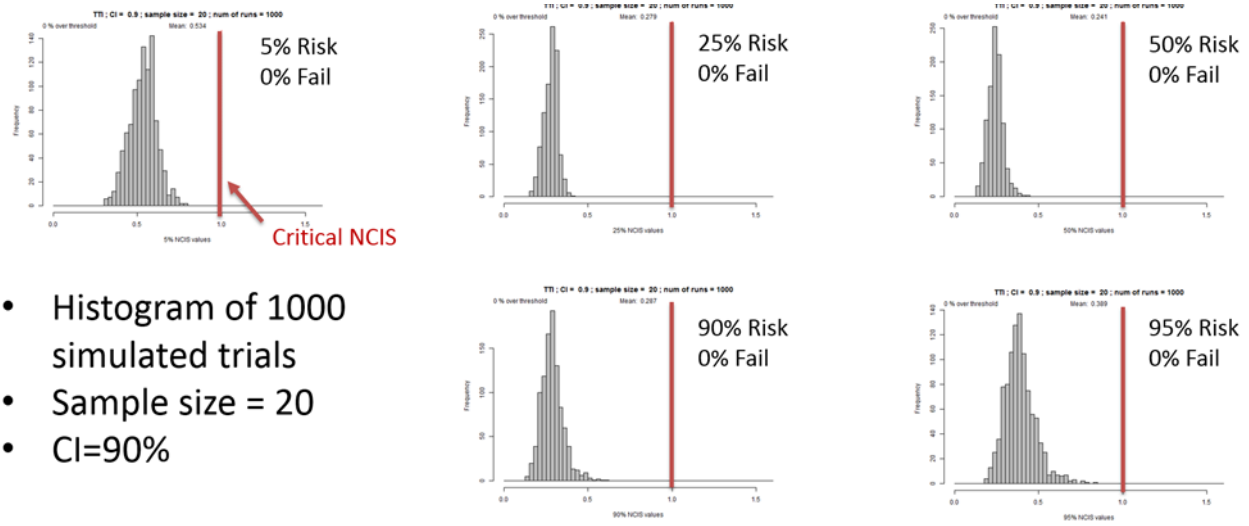


Figure A-4. TTI down-sampled from n = 30 to n = 20: red line is quality threshold for acceptable risk curve based on NCIS = 1.0; note, quality remains acceptable at all risk levels

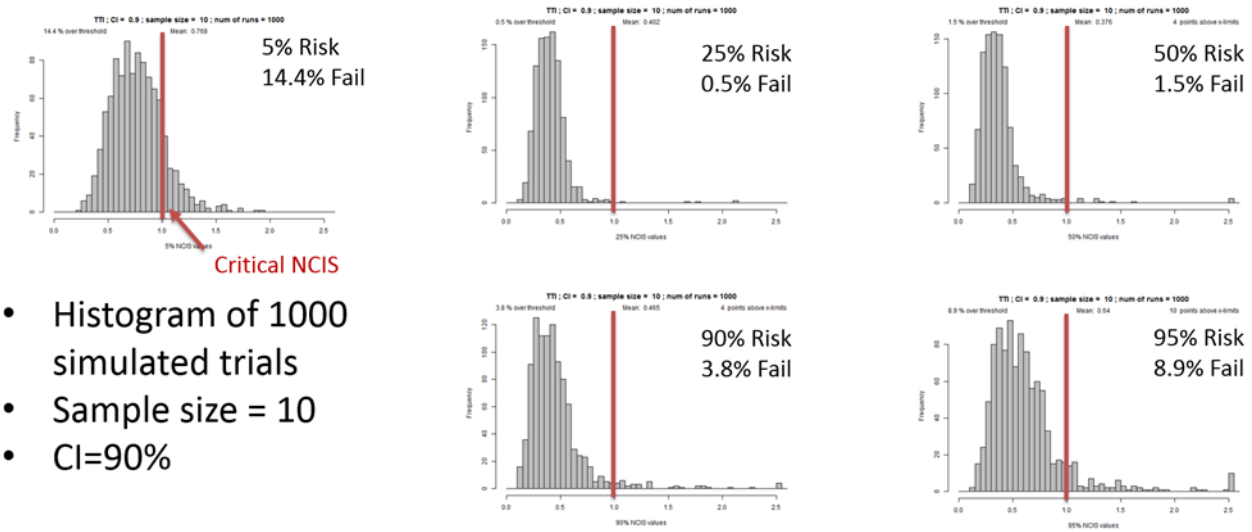


Figure A-5. TTI down-sampled from n = 30 to n = 10: red line indicates NCIS quality threshold of 1.0; note, chance exists of not meeting quality specifications at high and low risk levels

Even for the same data set, the degradation in quality can be more significant in other injury metrics. Figure A-6 shows for the same data set, lower spine acceleration (SPL), which passed preliminary quality inspection, that it has much worse NCIS values compared with TTI when down-sampled to n = 10. Here we see a net likelihood of failure to meet the NCIS targets at higher risk levels.

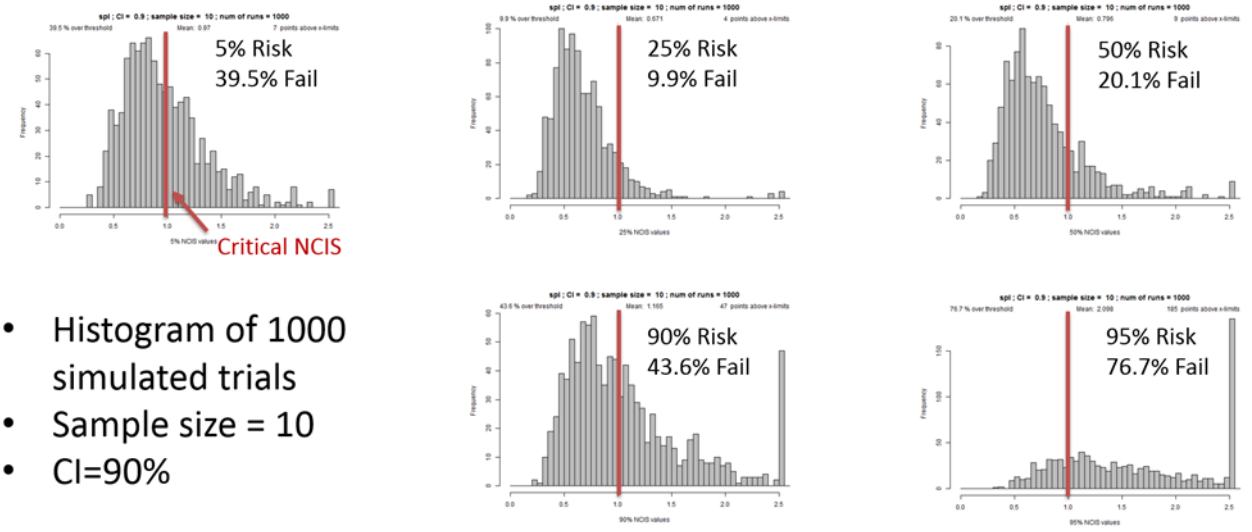


Figure A-6. SPL distribution of NCIS across different risk levels: red line indicates NCIS 1.0 threshold; note, significant number of resamples did not meet quality criteria, especially at the extreme risk levels (5% and 95%).

Appendix B – Determining Which Injury Metric from a Number of Potential Choices is Better

B.1 Scope

We seek to answer the question: How do you decide which injury metric is the better predictor? General guidelines are as follows:

- Take injury severity and mechanism into account.
- Consider practicality: limit human injury probability curves (HIPCs) to those that can be translated to the dummy.
- The Working Group (WG) can recommend the predictors by ranking each via the example above.
- The goal is to produce a clean data set for use in the HIPC code.

The aim of this section is to provide generic guidelines toward choosing injury metrics for modeling prediction of injury. Traditional statistical measures depend on the underlying model chosen and rely on asymptotic assumptions for p-values. A nonparametric measure is described, which works in general for any probabilistic binary decision model and evaluating the performance through extensive simulations. General pitfalls and deficiencies of the proposed measure are also described.

B.2 Methodology 1: Brier Method Score

B.2.1 Introduction

In survival analysis, there is some literature on mean-squared error, predictive performance, and measures on explained variation; however, it is mostly limited to right-censored data. None of the literature focused on choosing between time variables. The BMS is developed by Dr. Anjishnu Banerjee and colleagues. It is the first approach developed for evaluation of performance and explained variation in interval censored, left censored, and current status data.

B.2.2 BMS

For any postmortem human subject (PMHS) experiment, consider a set of n human cadavers on which crash impact tests are done. Suppose there are m possible biomechanical metrics that are obtained from the PMHS experiment. For each metric j and each sample i, we record the highest value of metric j for cadaver i that did not result in an injury R_i^j and the lowest value of the metric j for cadaver i that caused an injury L_i^j. The probability of injury happening at or before value of the metric f_j is

$$F(f_j) = 1 - S(f_j), \tag{B-1}$$

where $S(f_j)$ represents the probability of survival. The BMS (expected error rate) can be estimated by

$$e(\hat{f}_j) = \frac{1}{n}\sum_{i=1}^{n}(I_i(f_j) - \pi(f_j))^2, \quad \text{(B-2)}$$

where $e(\hat{f}_j)$ is the expected error rate, n is the number of human cadavers, $I_i(f_j)$ is the injury outcome of the experiment (either 1= injured or 0 = uninjured), and $\pi(f_j)$ is the probability of injury at or before value f_j for biomechanical metric j. The $I_i(f_j)$ may or may not be known depending on the censoring status. It will be estimated based on the censoring status (right censored, left censored, and interval censored) and calculated over data informative range (model independent) or over full range (model dependent). For the biomechanical metric j, we choose two extreme values, F_j^{min} and F_j^{max}, such that the range is a superset of the informative range of the values of the metric. The proposed BMS (i.e., cumulative error [*ce*] measure) will be estimated as follows. The evaluation of $e(\hat{f}_j)$ would require imputation of the unknown indicators. We use the following conditional expectation (E; the expected value of a random variable here) as a plug-in estimator for the unknown indicators. Beginning with Equation B-3 is the calculation for the model-independent approach.

For the unknown indicator corresponding to a right-censored observation,

$$ce = \frac{1}{R_j^i - F_j^{min}}\int_{F_j^{min}}^{R_i^j} e(\hat{f}_j)\, d(f_j), \quad \text{(B-3)}$$

where $e(\hat{f}_j)$ is replaced by $E(I_i(f_j)|R_i^j) = \frac{\pi(f_j)-\pi(R_i^j)}{1-\pi(R_i^j)}$ and $d(f_j)$is the differential of the variable f_j.

For the unknown indicator corresponding to a left-censored observation,

$$ce = \frac{1}{F_j^{max} - L_j^i}\int_{L_i^j}^{F_j^{max}} e(\hat{f}_j)\, d(f_j), \quad \text{(B-4)}$$

where
$$E(I_i(f_j)|L_i^j) = \frac{\pi(L_i^j)-\pi(f_j)}{\pi(L_i^j)}. \quad \text{(B-5)}$$

For the unknown indicator corresponding to an interval-censored observation,

$$ce = \frac{1}{F_j^{max} - L_j^i + R_j^i - F_j^{min}}\left\{\int_{F_j^{min}}^{R_i^j} e(\hat{f}_j)\, d(f_j) + \int_{L_i^j}^{F_j^{max}} e(\hat{f}_j)\, d(f_j)\right\}, \quad \text{(B-6)}$$

where $$E(I_i(f_j)|R_i^j, L_i^j) = \frac{\pi(f_j)-\pi(R_i^j)}{\pi(L_i^j)-\pi(R_i^j)}.$$ (B-7)

For the model-dependent approach, the formula is written as follows:

$$ce = \frac{1}{F_j^{max}-F_j^{min}} \int_{F_j^{min}}^{F_j^{max}} e(\hat{f}_j)\, d(f_j).$$ (B-8)

The biomechanical metric producing the lowest BMS is the most appropriate metric for further analysis.

B.3 Methodology 2: Receiver Operator Curve

B.3.1 Introduction

In the statistical modelling of an injury risk curve, it is important to decide which injury metric is the better predictor of injury among the available biomechanical metrics. Steps to evaluate rates of the two types of errors (false positives and false negatives) from any model that produces a binary classification probability are described. The methodology to construct an ROC, considering optimality of prediction rules and evaluation of injury metrics, is described.

B.3.2 Background

One of the first uses of ROCs was probably by the Allies during World War II, in the analysis of radar signals for detection of enemy objects (German planes). Thereafter, ROCs were used in signal detection (Green & Swets, 1966) and were applied in other areas, including psychophysics and medical technology (Green & Swets, 1966; Zweig & Campbell, 1993). In recent times, ROCs and their extensions have been extensively used in machine learning and statistical data mining for evaluation of prediction rules (Bi & Bennett, 2003).

B.3.3 Contingency Tables

One of the first steps in evaluation of a prediction rule is the construction of a contingency table based on a prespecified cutoff to look at specificity and sensitivity. Consider a logistic regression model for building an injury risk curve using hypothetical resultant upper spinal acceleration (RSPU) data as the injury metric. *After* fitting the model, a table of fitted probabilities for a range of RSPUs is obtained, as shown in Table B-1.

Table B-2. Logistic Regression for Hypothetical RSPU Data

RSPU Value	True Injury Status	Predicted Probability of Injury from a Logistic Model
55	No injury	0.01
60	No injury	0.04
65	No injury	0.13
70	Injury	0.34
75	No Injury	0.65
80	Injury	0.87
85	Injury	0.96

Based on the logistic regression model, any cutoff between [0, 1] is chosen for injury prediction: It is predicted that there will be no injury for any RSPU that leads to estimated injury probability less than or equal to the cutoff, and there will be injury otherwise. Choosing a cutoff value of 0.5, the following predictions are determined for the same hypothetical data set as in Table B-2.

Table B-3. Predicted Injuries Based on a Cutoff Probability of 0.5

True Injury Status	Predicted Probability of Injury from a Logistic Model	Predicted Injury Status
No injury	0.01	No injury
No injury	0.04	No injury
No injury	0.13	No injury
Injury	0.34	No injury
No Injury	0.65	Injury
Injury	0.87	Injury
Injury	0.96	Injury

Based on Table B-2, Table B-3 is constructed for evaluation of the injury metric and prediction rule (this is often called a contingency table or a confusion matrix). Table B-3 lists values of true positive rate (TPR) and true negative rate (TNR).

Table B-4. Contingency Table for Evaluation of Prediction

Predicted Injury Status	True Injury Status	
	Injury (TPR or sensitivity = 2/3)	TNR or specificity = 3/4
Injury	2	1
No Injury	1	3

B.3.4 Construction of the ROC

While the contingency table evaluates the accuracy based on a single cutoff for prediction, it would be of interest to evaluate the overall accuracy of the model based on a range of cutoffs.

Each cutoff produces estimated true and false positive rates. In the RSPU example, the cutoff value increases, and the TNR increases while the TPR decreases. For a cutoff of 0.96 or higher, the prediction would be all "no injuries", indicating a TNR of 100% and a TPR of 0%. In short, the ROC graphically shows the ability to discriminate between injury and noninjury. Metrics that keep the rank in order will generate an identical ROC.

An ROC graphically represents the TPR versus the false positive rate (FPR) for each possible cutoff. The basic layout of an ROC "space" is shown in Figure B-1. The red line represents an ROC that would be obtained with random guesses for injury and no-injury classification. The blue dots represent single-threshold examples of different classifiers (in our case metrics). Here A is the best threshold for a moderate classifier, B is not better than random, and C is the point of the best threshold for the best classifier within this group because of its location nearest the upper-left corner. The ROC is constructed by selecting a threshold, calculating TPR and FPR, and plotting that point on the axes (in Figure B-1). The curve would effectively be a series of these points. After a full sweep of thresholds, the ROC can be plotted in this space. The more an injury metric's ROC moves upward and leftward, the better the TPR and the lesser the FPR; hence, the better its predictive ability. The top-leftmost corner is the point with 100% TPR and 0% FPR (hence, 100% TNR), indicating perfect classification.

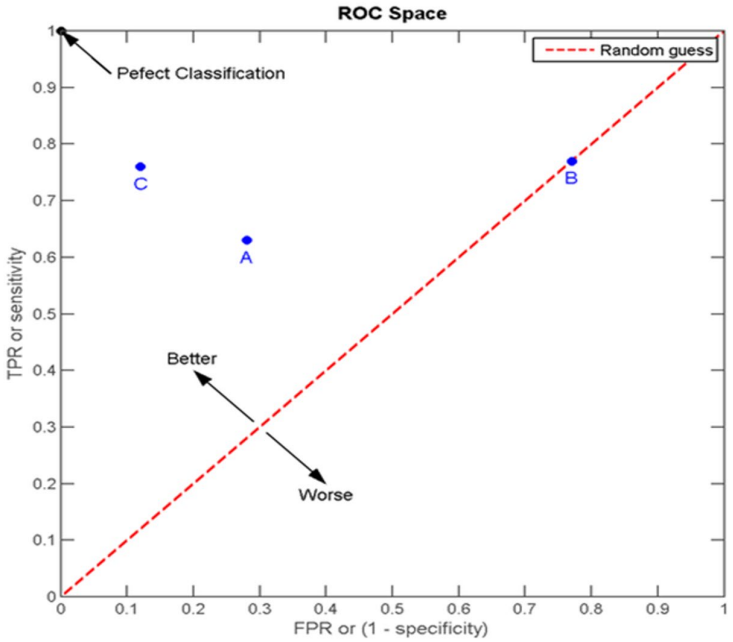


Figure B-7. Basic layout of an ROC

For the hypothetical RSPU data, the ROC based on Table B-3 is plotted in Figure B-2. The ROC, represented by the red line, is a step function in this simple case, with jumps at the places cutoffs would represent differences in the TPR and FPRs. It is worthwhile to note that ROC can

be constructed for any probabilistic binary classification model. The logistic regression model is used here for demonstration purposes only.

B.3.5 Area under the ROC

The AROC is a single measure of the overall predictive accuracy of the model. The perfect ROC has an area of 1, while an ROC with an area of 0.5 indicates the model is no better at predicting than random guesses. The AROC is typically calculated by segmenting the curve into trapezoids, calculating the area for each one, and then summing them up. In our example with hypothetical RSPU values, the area under the curve is 0.92 (Figure B-8). After appropriate normalization, the AROC value is the probability the estimate for a randomly chosen injury sample point is higher than a randomly chosen no-injury sample point. Given two injury metrics from the *same underlying sample*, two statistical models can be fit that estimate the probability of injury for each injury metric. From the fitted values, the AROC is computed for each model. *The metric with the higher AROC value is the better overall predictor of injury*.

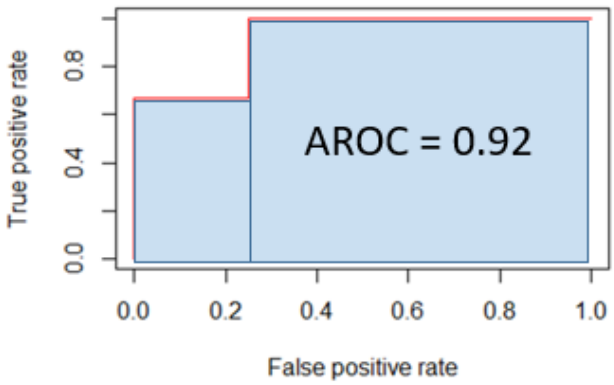


Figure B-9. ROC for hypothetical RSPU data

B.3.6 Summarization of Procedure to Compare Injury Metrics

The following steps summarize the methodology for comparison of two injury metrics, say A and B.

B.4 Algorithm

1. Obtain a sample containing both injury and no-injury points and the values of injury metrics A and B for each point. If the sample contains only injury points or only no-injury points, classification is meaningless because any cutoff would give perfect classification.

2. Fit a probabilistic model for predicting injury using metrics A and B independently. WG research has shown that the AROC is a function of the data and will not be affected by which survival analysis was chosen.

3. Obtain the AROC for each of the fitted models.

4. If both AROCs are below 0.5, it is meaningless to compare them because none of the models has any predictive ability. If at least one model has an AROC greater than 0.5, the one with higher AROC has better predictive ability.

It is trivial to extend this algorithm for comparison of more than two metrics.

If a sufficiently large sample is available, a fraction of the sample can be kept outside the initial data pool when the model is fit (called the held-out sample). Predictive values are then obtained by evaluating the fitted model on this held-out sample and corresponding AROCs can be compared. This process is called cross-validation and may reduce over-fitting, the discussion of which is beyond the scope of this article, but references can be found in Bi and Bennett (2003) and Picard and Cook (1984). A rule of thumb sometimes used: When the sample size is at least 50, use at least 20% of the sample as the held-out fraction, while ensuring that the held-out fraction has both injury and no-injury values.

B.4.1 Discussion

There are several *advantages* of using the AROC to choose between injury metrics. The AROC is a nonparametric measure, free of distributional or model assumptions. Statistical measures like the Wald's p-value rely on asymptotic approximations. The AROC as a measure is free from approximations and is a function only of the predictive ability of the model. It is not meaningful to compare Wald's or similar statistics across different models; for example, one cannot meaningfully compare between a probit and a logit regression model using the Wald's p-values obtained separately from each of them. When using the same underlying sample, it is meaningful to compare between methods and different models, using the AROC, for evaluating predictive ability.

Some *limitations* of the proposed measure exist. AROC needs to be used with caution when interval censored data are available. It is not meaningful to compare AROC across different samples. There must be a mix of injury and no-injury points in the sample for the AROC to be valid. Any sample having only injury or only no-injury points will always have an AROC of 1, irrespective of the fitted probabilities. Finally, AROC can be noisy and unreliable if the sample does not have a good mix of no-injury and injury values. For example, a sample consisting of 19 injury points and 1 no-injury point will not produce reliable comparison between injury

metrics using AROC. It is worthwhile to note that logistic regression or similar techniques in these scenarios will also produce unreliable estimates, with high uncertainty/variance.

Appendix C – Review of Human Injury Probability Curve (HIPC) Working Group (WG) Efforts to Verify Performance of Brier Method Score (BMS)

Several studies were performed within the WG to evaluate the performance of the BMS. They were presented from October to November 2016. These studies included simulated data sets that showed the performance of the metric-selection methods for different types of censoring.

Two different BMS (model dependent and model independent) are shown here. The method used throughout this report and the HIPC code is the Model Independent version. The difference between the model dependent and model independent is that the model dependent assumes a linear risk curve (i.e., imputes data) in the nonobserved regions of a censored observation (Figure C-1) when evaluating the fit against that observation point. The model independent does not evaluate the fit against nonobserved regions (Figure C-2). In both Figure C-1 and Figure C-2, the min and max are the range of observed criteria.

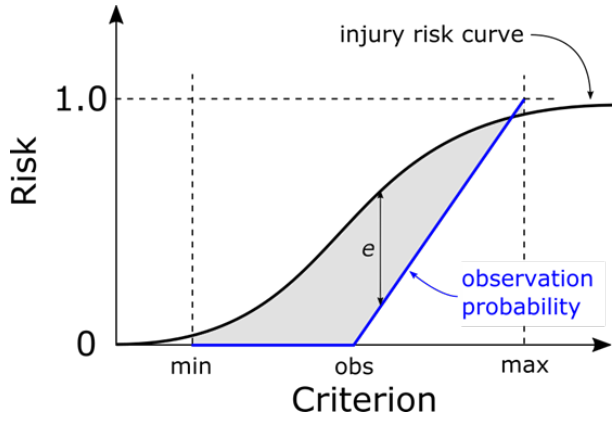


Figure C-10. Evaluation region for right-censored data point using model-dependent BMS; although region greater than the observation (obs) is not directly observed, injury risk is still imputed for this region using linear assumption

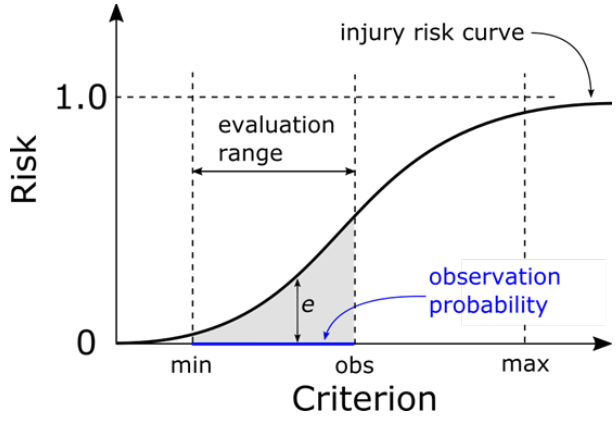


Figure C-11. Evaluation region for right-censored data point using model-independent BMS; note, region greater than the observation (obs) is not used to evaluate the curve's fit

Figure C-3 shows the distributions used to generate simulated data sets. In general, a better metric (blue) is one that better discriminates between low risk of injury and high risk of injury. It has a higher slope between the noninjury level of the criteria to an injury level. Figure C-4 and

Figure C-5 show examples of simulated results and how many times the method chose the data set generated using the best injury-risk curve. For all simulated data, we found the model-independent approach generally picked the best distribution. For this reason, this approach was adopted for the HIPC code. In these cases, combined, doubly and interval censored data refer to the type of observation simulated. These simulated observations were generated by simulating a value of injury and also a value (or values for interval censoring) to make observations. If the observation is above the injury point, an injury would be observed at the observation value. A visual representation of this is found in Figure C-6 for a simulated right-censored observation. This is similar to the method used by Cutcliffe et al. (2012).

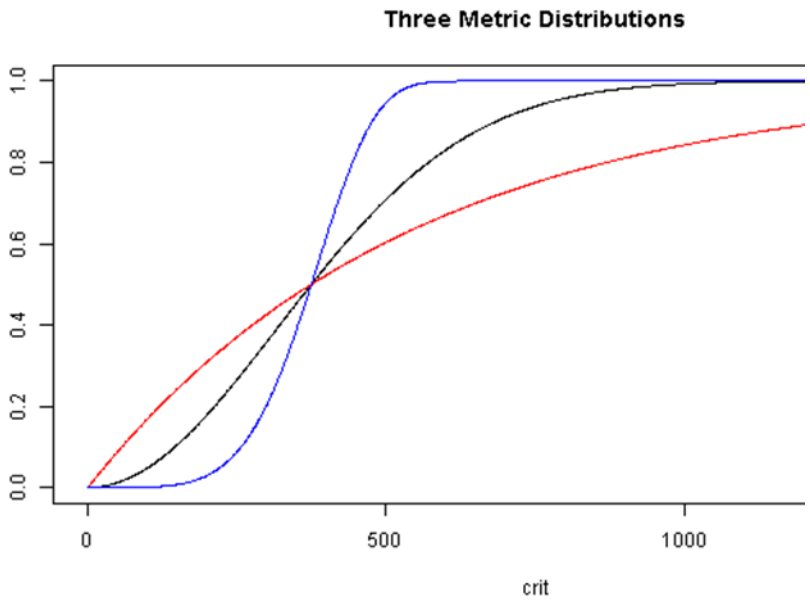


Figure C-12. Simulated distributions (metrics) based on Weibull distributions used to generate simulated data sets; blue curve would be considered the best followed by black, and red curve would be a poor metric because it poorly discriminates injury risk

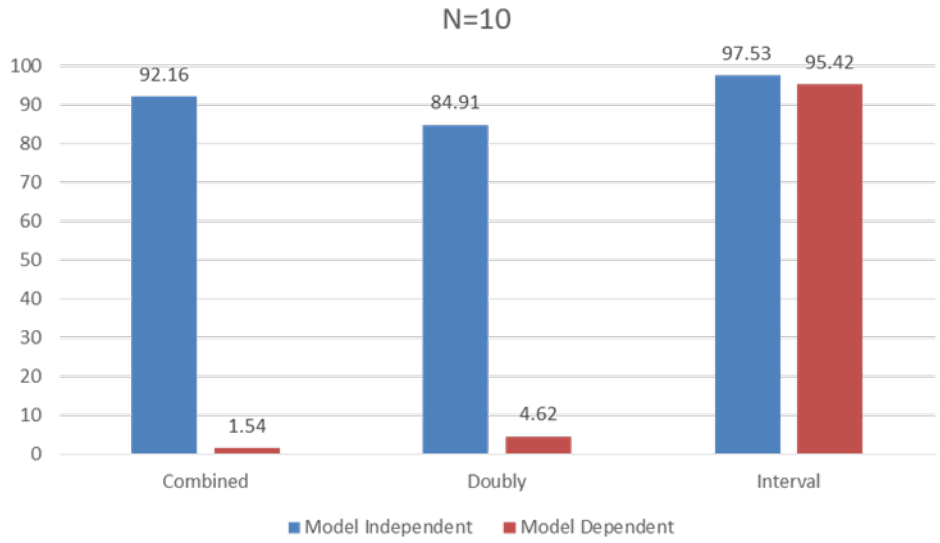


Figure C-13. Performance of BMS (recommended approach—model independent) for various types of censoring; 10 samples simulated yielding 10 data points

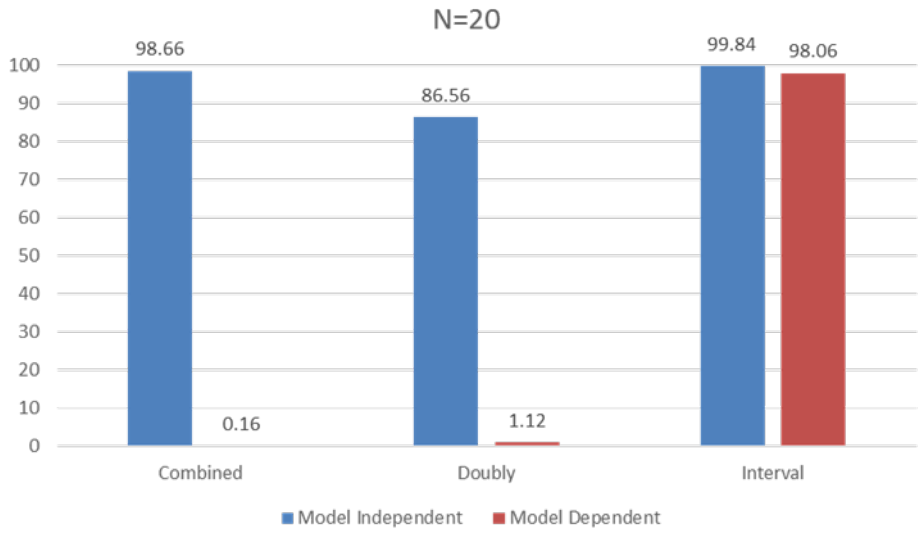


Figure C-14. Performance of BMS (recommended approach—model independent) for various types of censoring; 20 samples simulated yielding 20 data points

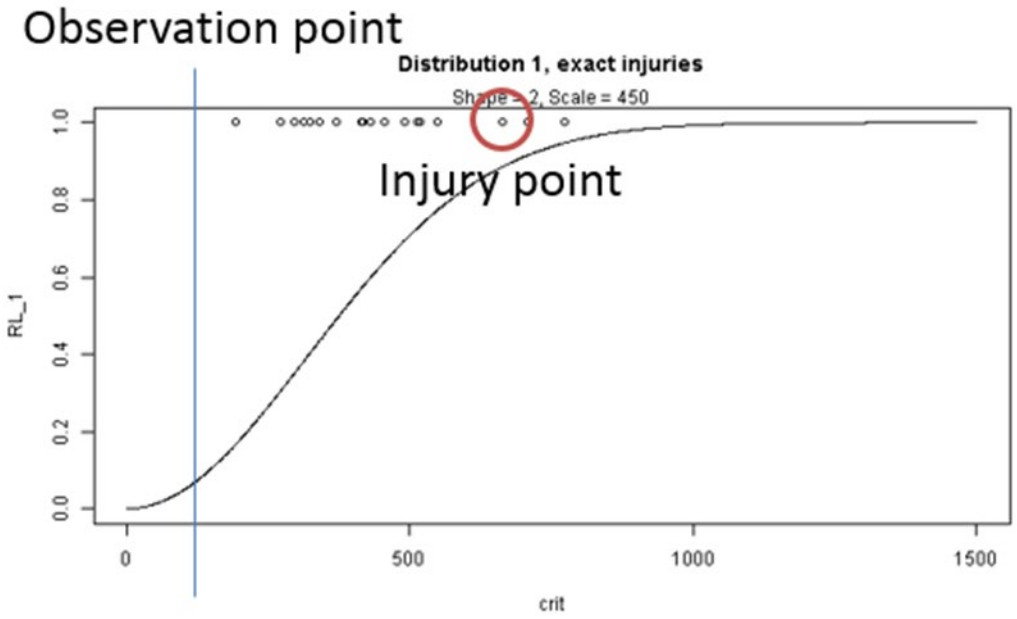


Figure C-15. How data were simulated when generating data sets for validating BMS: here, injury point (circled in red) is higher than observation point (vertical blue line), and this becomes right-censored (noninjury) observation at blue line

Appendix D – Glossary

Anthropomorphic test device (ATD): A physical device having the size, shape, and biomechanical response approximating that of a live human. ATDs used in dynamic testing also have instrumentation that measures mechanical parameters that have associated Injury Assessment Reference Curves (IARCs) and thus can be used to predict the potential for particular types of injuries.

Data censoring: Data censoring is employed when the exact point of injury for a specimen is unknown. Interval censoring requires that a specimen is tested under both noninjury and injury conditions and uses the peak value of the injury metric under noninjury and injury conditions as the lower- and upper-bound data for that specimen, respectively. Right censoring is used for specimens that do not have injury. Left censoring is used for specimens that only have injury test data.

False negative: In case of a binary decision rule, it is the rule predicting there was no event when an event occurred. In this case, it is the prediction of no injury when an injury did happen.

False negative rate (FNR): Defined as (1-true positive rate) or (1-sensitivity).

False positive: In case of a binary decision rule, it is the rule predicting an event in the absence of an event. In this case, a false positive is the prediction of an injury when there was none.

False positive rate (FPR): Defined as (1-True Negative Rate) or (1-specificity).

Human injury probability curve (HIPC): A statistical relationship between a continuous mechanical parameter measured in physical tests (or calculated from the results of physical tests) with a postmortem human surrogate and the probability of a particular injury.

Injury Assessment Reference Curve (IARC): A mathematical relationship between a continuous mechanical parameter measured (or calculated from the results of physical tests) in physical tests or simulations with an ATD and the probability of a particular injury to a human.

Injury Risk Curve (or risk curve): A generic term for an HIPC.

Normalized Confidence Interval Size (NCIS): A mathematical representation of the width of a confidence interval divided by the independent variable, at a given risk level. (See Equation 1.) Also known as the Quality Index in the International Organization for Standardization (ISO) injury risk curve protocol.

Postmortem human subject (PMHS): A term for cadavers used in biomechanical tests that recognizes cadavers are only partially representative of live humans.

Reproducibility: Metric of correlation between a set of tests and a single test removed from the set.

Sensitivity: The probability of a model or a decision rule to have a true positive at a randomly chosen event (in this case injury) sample point. This is often estimated as the proportion of true positives to the number of events from a model or decision rule applied to a data set. This is also called the TPR or recall.

Specificity: The probability of a model or a decision rule to have a true negative at a randomly chosen nonevent (in this case, no-injury) sample point. This is often estimated as the proportion of true negatives to the number of nonevents from a model or decision rule applied to a data set. This is also called the TNR.

True negative: The opposite of a false negative—correctly predicting no injury when no injury occurred.

True positive: The opposite of a false positive—correctly predicting an injury when it actually occurred.

Warrior Injury Assessment Manikin (WIAMan): Army-sponsored program to develop an ATD specifically designed to predict injury risk in vertical-loading environments.

WG: Working Group.

Appendix E – List of Acronyms

AFT	Accelerated Failure Time
AIC	Akaike information criterion
AROC	area under the receiver operator curve
ATD	anthropomorphic test device
BMD	bone mineral density
BMS	Brier Method Score
CI	confidence interval
FNR	false negative rate
FPR	false positive rate
HIPC	human injury probability curve
IARC	Injury Assessment Reference Curve
IARV	Injury Risk Reference Value
ISO	International Organization for Standardization
KS	Kolmogorov–Smirnov
MCW	Medical College of Wisconsin
NCIS	normalized confidence interval size
PCA	principal component analysis
PH	Proportional Hazards
PMHS	postmortem human subject
QQ	quantile–quantile
RC	risk curve
RCG	Risk Curve Generator
ROC	receiver operator curve
RSPU	resultant upper spinal acceleration
SW	software
TNR	true negative rate
TPR	true positive rate

TTI	Thoracic Trauma Index
WG	Working Group
WIAMan	Warrior Injury Assessment Manikin
WorldSID	Worldwide harmonized Side Impact Dummy;

Appendix F – Distribution List

ORGANIZATION

U.S. Army CCDC Data & Analysis Center
FCDD-DAS-LBG/T Resetar-Racine
6908 Mauchly Street
Aberdeen Proving Ground, MD 21005

U.S. Army CCDC Data & Analysis Center
FCDD-DAS-LBW/ K. Loftis
FCDD-DAS-LBW/D. Barnes
6904 Magazine Road
Aberdeen Proving Ground, MD 21005

U.S. Army CCDC Army Research Laboratory
FCDD-RLD-CL/Tech Library
2800 Powder Mill Rd.
Adelphi, MD 20783

Defense Technical Information Center
ATTN: DTIC-O
8725 John J. Kingman Rd.
Fort Belvoir, VA 22060-6218