EVALUATING PERFORMANCE OF SURVIVAL REGRESSION MODELS WITH INTERVAL CENSORED DATA IN MOTORVEHICLE CRASH EXPERIMENTS

By Anjishnu Banerjee[*], Narayan Yoganandan[*], Fang-Chi Hsu[†], Scott Gayzik[†] and Frank A. Pintar[‡]

Medical College of Wisconsin [*], *Wake Forest University* [†] *and VA Medical Center* [‡]

The presence or absence of injures identified from testing human cadavers (termed post mortem human subjects, PMHS, in impact biomechanics literature) are used in conjunction with biomechanical outcomes such as force, deflection, and acceleration to derive injury probability curves. Injury probability curves are crucial in the design and improvement of saftey interventions such as seat belts and airbags. According to latest ISO recommendations, survival analysis has been suggested for the estimation of these injury probability curves. However, for survival models, we need to choose between multiple biomechanical metrics, which are routinely collected as apart of PMHS experiments. We analyze performance measures for survival regression models in these contexts to help discriminate between the metrics. We propose a new class of measures, explore their theoretical properties and extensively assess them in a variety of simulated data scenario and a PMHS test dataset. Our proposed class of measures have applicability in generic right, censored , interval or current status data settings.

1. Introduction.

Road traffic crashes are among the leading causes of injury/death. According to the Centers for Disease Control and Prevention (CDC), the lifetime medical cost of motor vehicle crash injuries in 2012 was US dollars 18 billion and more than 75% of this was attributed to incur during the first 18 months following the injury. The lifetime work lost due to these injuries is estimated to 33 billion, while the total value of societal harm was estimated as US dollars 836 billion from motor vehicle crashes alone (Blincoe et al., 2015). CDC also stated that the “best way to keep people safe and reduce medical costs is to prevent crashes from happening in the first place. But if a crash does occur, many injuries can still be avoided through the use of proven interventions” One of the proven interventions is continued vehicular improvements in the design of safety measures such as sensors, airbags, seatbelts and load limiters to limit the impact force applied

Keywords and phrases: Motor Vehicle Crash Tests, Performance Evaluations

on any occupant in the automobile, and other vehicle interior and exterior structures. This is accomplished through impact biomechanics research.

Injury risk curves form the primary basis for mitigation of injuries and fatalities in environments such as motor vehicle crashes (Kleinberger et al., 1998) and events such as underbody blast (UBB) incidents from Improvised Explosive Devices(IEDs)(Danelson et al., 2015). Risk curves are used by the United States Federal Motor Vehicle Safety Standards(FMVSS) for evaluating crashworthiness and safety of vehicles sold in the USA, consumer information appearing as star ratings in the MSRP stickers displayed on the automobiles for public awareness, and designs of safety systems such as seatbelts and airbags.

Injury risk curves are typically estimated from data obtained from testing human cadavers, often called post mortem human subject (PMHS) experiments (Nahum and Melvin, 2012). After the completion of a PMHS experiment, for each subject under observation, data for a variety of biomechanical metrics and injury status corresponding to these biomechanical metrics are collected, sometimes accompanied by other subject specific variables of interest, such as age and sex, bone mineral density, body mass index, among others. Biomechanical metrics could include directly observed ones from the experiment such as peak force, deflection, or, metrics derived from combinations of other metrics/parameters, such as combined accelerations from different levels of the rib/spine. Injury outcomes are dichotomized into presence or absence of injury based on observations of severity of fractures, organ trauma and other biological measures. Development of an injury risk curve is then done by estimation of the risk corresponding to each value of a biomechanical metric, such as peak force, while adjusting for observed demographic covariates, if present and other experimental conditions.

Traditionally, risk curves have been developed using binary logistic regression, estimating the probability of injury, given the observed values of the biomechanical metrics and other covariates (Robbins, Melvin and Stalnaker, 1976; Kuppa et al., 2003; Philippens et al., 2009). However, more recently, survival analysis techniques have been used for risk curve estimation in these impact biomechanic contexts (Kent and Funk, 2004; Petitjean et al., 2009), where presence or absence of injury correspond to events and the biomechanical metric in question corresponds to "time", so that instead of "time-to-event" data in traditional survival analysis, we have "force-to-injury" or "acceleration-to-injury" type data in the present context. To better elucidate this setting, let us consider a toy example, when two samples from a PMHS test have yielded a peak force of 10 and 20 Newtons, with the 10 Newton force resulting in no-injury, while the 20 Newton force results in injury. Even though the 20 Newton force resulted injury, this is not necessarily the exact injury causing force - an injury could well have occured for a force lesser than 20 Newtons. It is then reasonable to assume for modeling purposes that the 20 Newton force represents a "left-censored" observation in survival analysis, while following a similar argument, the 10 Newton force represents a "right-censored" observation. Accounting for censoring information through survival analysis yields better risk profiles than logistic regression - the ISO/TC22/SC12/WG6 working group of the International Standards Organization (ISO) now recommends using survival analysis over binary logistic for the development of risk curves.

There are however several challenges in adapting survival analysis techniques for risk curve estimation in impact biomechanics contexts. A majority of survival analysis techniques consider uncensored and right censored data, while the PMHS experiments only yield right, left or interval censored (when more than one test has been performed on a single sample) data, with no uncensored observations, bearing conceptual similarity to observations in current status data. Additionally, survival analysis, typically accommodates only one "time" variable, while PMHS experiments commonly record multiple biomechanical metrics, most of which could serve as the "time" equivalent. There is very limited previous work in looking at multiple candidate "time" variables. In a PhD thesis Duchesne (1999), considers multiple time-scales, such as time since purchase and mileage in studying automotive reliability, where each of the time like measures are a function of the actual time variable. Similar ideas are echoed in Oakes (1995); Kom, Graubard and Midthune (1997). However, in our context, we do not have a underlying actual "time" that all the recorded biomechanical metrics are a function of. In this paper, we focus on developing a rigorous statistical frame-work for defining valid "time-type" metrics, developing performance measures and analyzing the performance of each of metrics using these measures to choose the metric best suited for the development of risk curves.

Our contribution in this article is two fold. While there is some literature on mean-squared error, predictive performance and measures on explained variation in survival analysis, it is exclusively limited to right censored data typically arising in biomedical contexts (Graf et al., 1999; Gerds and Schumacher, 2007; Gerds, Cai and Schumacher, 2008), none of the previous focused on or relevant to choosing between time variables. To the best of the authors' knowledge, the measures proposed in this article are the first methods developed for systematic evaluation of performance and explained variation in generic interval censored, left censored and current status data. Secondly, to the best of the author's knowledge, this is the first statistical performance evaluation of risk curves for impact biomechanics research, where previously ad-hoc measures of quality have been considered and no measures for explained variation exist for choosing between biomechanical metrics. We believe that improved risk curves using our proposed methods would significantly advance the state-of-the-art in currently held standards for motor vehicle crashworthiness. We also believe that our methods have applications beyond impact biomechanics research, being applicable to other contexts in current status data and competing "time" variables.

The rest of the paper is organized as follows: In section 2, we begin with a description of basic experimental set-up for PMHS crash tests and some notations that are helpful throughout the paper. In section 3, we consider performance measures and in section 4, we evaluate different scenarios with simulated data. In section 5, we consider a set of PMHS experiments, with 24 biomechanical metrics, for which only logistic regression has been done before. We end with final conclusions and discussion in section 6. Technical details and proofs are provided in the supplementary material.

2. Set up and notational preliminaries.

2.1. *Data descriptives.* For any PMHS experiment, consider a set of n human cadavers on which crash impact tests are done, indexed by $i = 1, 2, \ldots, n$. In some recent experiments, development of new sensing methodology such as strain gages and accoustic emission sensors (Gallenberger, Yoganandan and Pintar, 2013), might lead to methodology to obtain uncensored "force-to-event" type data for PMHS experiments. The biomechanical community is however yet to develop consensus on the use and processing of signals from these sensors. In light of this, we shall only consider observations which are either right, left or interval censored.

Suppose that there are m possible biomechanical metrics that are obtained (observed or derived) from the PMHS experiment, indexed by $j = 1, 2, \ldots, m$. For each such metric j and each sample i, we record the tuple (R_i^j, L_i^j), which is obtained as follows. R_i^j is the highest value of metric j for cadaver i which did not result in an injury. L_i^j is the lowest value of the metric j which caused an injury. In case of cadavers with multiple crash tests, none of which resulted in an injury, we shall only have a value for R_i^j in the tuple and will use the notation 'NA' for L_i^j, representing a right censored observation. In case of a cadaver with a single crash test which resulted in an injury, we only have a value for L_i^j, with 'NA' being recorded for R_i^j, representing a left censored observation. In case of a cadaver with multiple crash tests, some of which did not result in injuries, while at least one test resulted injury, we shall obtain values for both elements of the tuple

(R_i^j, L_i^j), representing an interval censored observation. Note that values of a metric j, which are $< R_i^j$, which did not cause an injury for cadaver i, are non-informative and are not recorded. Corresponding to the recorded tuple (R_i^j, L_i^j), we consider a censor status indicator variable, Δ_i^j, which will take the values $0, 1, 2$ corresponding to a right, left and interval censored observations respectively. For a cadaver i, $Z_i = \{Z_i^k, \text{for } k = 1, 2, \ldots, p\}$ shall denote the covariates specific to i. For a cadaver i and metric j, T_i^j represent the unobserved and uncensored value of the metric, the "exact" breaking force or biomechanical metric at which an injury would have happened. Specific to cadaver i and metric j, we shall denote the probability of injury happening at or before value of the metric f_j as,

$$F(f_j) = \Pr(T_i^j \leq f_j) = 1 - S(f_j),$$

where $S(f_j)$ represents the probability of survival, that is, $S(f_j) = \Pr(T_i^j > f_j)$. We have intentionally dropped covariates Z_i from the expressions for $S(\cdot)$ and $F(\cdot)$, for notationally clarity, whenever covariates are present, the dependence shall be assumed inherently. Finally, let $I(f_j)$ and $I_i(f_j)$ be indicators of injury at or before the value f_j for the metric j, with $I(\cdot)$ denoting the generic random variable and $I_i(\cdot)$, the observation corresponding to sample i. Therefore, $I(f_j) = 1$ [or for sample specific observations, $I_i(f_j) = 1$] if $T^j \leq f_j$ [for sample specific observations if, $T_i^j \leq f_j$] and $= 0$ otherwise. We shall use $\mathbb{E}(X)$ to denote the expected value of a random variable X.

2.2. *Valid metrics and observations.* Not all biomechanical metrics are always available or perfectly recorded for all cadaver tests, due to limitations of testing procedures. We impose a few restrictions on the values of the observed tuples, so that they lead to valid estimation algorithms for risk of injury. For any biomechanical metric to be considered as a candidate metric in our evaluation procedure, it cannot have solely right censored or solely left censored observations, since these observations do not lead to valid survival estimates (Gentleman and Geyer, 1994; Fay and Shaw, 2010). Therefore a metric j, for which either all R_i^j's are 'NA' or all L_i^j's are 'NA', shall be dropped from the analysis as not being a valid candidate metric. In some rare occasions, due to limitations of the experimentation environment, it may happen that an interval censored tuple (R_i^j, L_i^j) is incorrectly recorded such that $R_i^j > L_i^j$, in such a case, if a value, f_j of the metric j is available for a test on the cadaver i such that $f_j < R_i^j < L_j^i$, the value of R_i^j shall be replaced with f_j inspite of this not being the largest non-injury value; in case such a test does not exist, the observed tuple is converted to a left censored observation with the recorded value of R_i^j replaced with 'NA' and $\Delta_i^j = 2$.

3. Estimation and performance evaluation.

3.1. *Survival probability estimation.* In the presence of multiple candidates for time, one approach for survival analysis might to consider estimation of joint probabilities such as,

$$F(f_1, f_2, \ldots, f_m) = \Pr(T_i^1 \leq f_1, T_i^2 \leq f_2, \ldots, T_i^m \leq f_m),$$

which would be conceptually similar to methods in Berzuini and Clayton (1994). Even though this is theoretically possible, this is made practically infeasible due to the sample size in typical PMHS experiments with typical sample sizes in the range 10 to 30. In the real data example considered in this paper, we have 30 samples, with 24 metrics being recorded - therefore estimation of injury probabilities on a multivariate grid of all time like biomechanical metrics being infeasible. In what follows, we consider instead modeling of each biomechanical metric at a time.

Injury curves developed in the biomechanical literature typically rely on parametric survival regression, such as survival analysis using a underlying Weibull distribution assumption for the survival times(Kent and Funk, 2004; Petitjean et al., 2009). While Kaplan Meier and other similar non-parametric methods for survival models have been extended to accommodate interval censored data (Turnbull, 1976; Gentleman and Geyer, 1994; Sun, 1996), non parametric estimation yield step-functions for estimated survival probabilities, which are not desirable in the present context of injury curve estimation. An exception is in the case of perfectly or quasi separated observations, when for a biomechanical metric j we have,

$$\max_{i=1,2,\ldots,n} R_i^j \leq \min_{i=1,2,\ldots,n} L_i^j,$$

parametric survival regressions estimates may not converge and we have to resort to non parametric methods for the metric j. It is worth noting that there are many ways of survival probability estimation for current status data and interval censored data, including several recent developments, (Zhao et al., 2015; Ma, Hu and Sun, 2015; Chen et al., 2014), many of which may be adapted to our context. However our focus in this paper is not in this estimation process, but rather on development of performance measures for distinguishing between biomechanical metrics for any particular estimation process - and the measures we propose will be generally applicable to any valid survival probability estimation procedure.

3.2. *AUC type measures.* Area under the receiver operating characteristic curve (AuROC) has been used for binary classification models to provide a single measure of predictive ability of the model (Swets, 2014; Jiménez-Valverde, 2012). Using estimated survival probabilities, it is trivial to construct an ROC curve, treating presence/absence of injury as our classes of interest. It could also be used as a predictive performance measure for comparing the different biomechanical metrics of interest. There are however several issues which limit its applicability. The AuROC, in the context of survival analysis, does not take into account censoring, which can lead to misleading performance evaluations. Time varying extensions of the AuROC have been proposed to better deal with censoring (Blanche, Dartigues and Jacqmin-Gadda, 2013; Gerds et al., 2013), but these too clearly suffer from many deficiencies. The AuROC, which can be shown to be equivalent to the Mann Whitney U statistic (Hanley and McNeil, 1982), as well as its time varying counterparts are rank based measures, solely dealing with classification and not calibration. To elucidate this, consider again a toy example, with two biomechanical metrics yielding survival probability estimates $\{0.2, 0.3, 0.4, 0.5\}$ and $\{0.01, 0.49, 0.51, 0.9\}$ from some estimation model for 4 samples in a PMHS experiment, with injury status, $\{0, 1, 0, 1\}$ respectively. Clearly the probability estimates from the second metric provides provide better calibration, however the AuROC measures between the two would be the same. Several other issues exist with AuROC - they are known to not to be proper scores and can lead to misleading classification (Mol et al., 2005; Lobo, Jiménez-Valverde and Real, 2008), limiting their applicability. In our simulation section, we provide further examples of the limitations of AuROC measures in our contexts.

3.3. *Measures of explained variation for survival data at a fixed "time".* Let us denote by $\pi(f_j)$ an estimator for $F(f_j)$, the probability of injury at or before value f_j for biomechanical metric j, using some survival probability estimation procedure [as mentioned before, we drop the explicit dependence on covariates for notational clarity, the dependence will be assumed to be present whenever covariates are recorded and relevant]. Then, the expected error rate $e(\cdot)$ at f_j is given by,

$$e(f_j) = \mathbb{E}(I(f_j) - \pi(f_j))^2. \tag{1}$$

This is equivalent to the squared error loss or the Brier score (Gerds and Schumacher, 2007; Gerds, Cai and Schumacher, 2008). With n samples, this error rate in equation (1) can be estimated by,

$$e(\hat{f_j}) = \frac{1}{n}\sum_{i=1}^{n}(I_i(f_j) - \pi(f_j))^2. \tag{2}$$

In the presence of censoring, Gerds and Schumacher (2007) recommend using inverse probability of censoring weights as weights in equation (2) to adjust for potential censoring bias. We note that, in our context, whence all observations are censored, such adjustment is not necessary.

In equation (2), the quantity $I_i(f_j)$ may or may not be known depending on the censoring status. If the observation for metric j for sample i is a right censored observation, yielding the tuple, $(R_i^j, `NA')$, one of the following two scenarios occur,

1. Either $f_j \leq R_i^j$, in which case, $I_i(f_j) = 0$.
2. Or, $f_j > R_i^j$, in which case, $I_i(f_j)$ is unknown.

Similarly for a left censored observation, $(`NA', L_i^j)$, one of the following two scenarios occur,

1. Either $f_j \geq L_i^j$, in which case, $I_i(f_j) = 1$.
2. Or, $f_j < L_i^j$, in which case, $I_i(f_j)$ is unknown.

Finally for an interval censored observation, (R_i^j, L_i^j), one of the following three scenarios occur,

1. Either $f_j \leq R_i^j$, in which case, $I_i(f_j) = 0$.
2. Or, $f_j \geq L_i^j$, in which case, $I_i(f_j) = 1$.
3. Or, $R_i^j < f_j < L_i^j$, in which case, $I_i(f_j)$ is unknown.

3.4. *Error evaluation over "time" ranges.* The procedure discussed so far yields a measure specific to single value f_j for metric j. To obtain a general measure, over the range of values we consider two possible options.

3.4.1. *Evaluation over full range.* For the biomechanical metric j, we choose two extreme values, F_j^{min} and F_j^{max}, such that the range $(F_j^{\text{min}}, F_j^{\text{max}})$ is a superset of the likely range of the values of the metric that could have been observed in the course of the experiment. In the absence of user specified reasonable guesses, one could also invert the 1^{th} and 99^{th} quantiles of the fitted distribution $\pi(\cdot)$ to obtain reasonable boundaries, F_j^{min} and F_j^{max} or simply use the minimum and maximum of observed values. F_j^{min} could be chosen as 0, assuming that the biomechanical metric takes only non-negative values - however extreme tail probability estimates can sometimes be distorted, leading to inflated error metrics. If the full range route is followed, we recommend choosing quantile inversion based boundaries - the same quantiles chosen for all metrics keeps all the evaluations on a similar footing. Once the range is choose, a cumulative error measure could for each metric j may be obtained as,

$$ce = \frac{1}{F_j^{\max} - F_j^{\min}} \int_{F_j^{\min}}^{F_j^{\max}} e(\hat{f_j}) d(f_j). \tag{3}$$

However, evaluation of $e(\hat{f_j})$ would require imputation of the unknown indicators. We use the following plug-in estimators for the unknown indicators may be obtained from their conditional expectation based on the censoring status, namely,

1. For the unknown indicator corresponding to a right censored observation, and with $f_j > R_i^j$,

$$\mathbb{E}(I_i(f_j)|R_i^j) = \Pr(T_i \leq f_j | T_i > R_i^j) = \frac{\pi(f_j) - \pi(R_i^j)}{1 - \pi(R_i^j)}$$

2. For the unknown indicator corresponding to a left censored observation, and with $f_j < L_i^j$,

$$\mathbb{E}(I_i(f_j)|L_i^j) = \Pr(T_i \leq f_j | T_i \leq L_i^j) = \frac{\pi(f_j)}{\pi(L_i^j)}$$

3. For the unknown indicator corresponding to an interval censored observation, and with $R_i^j < f_j < L_i^j$,

$$\mathbb{E}(I_i(f_j)|R_i^j, L_i^j) = \Pr(T_i \leq f_j | R_i^j \leq T_i < L_i^j) = \frac{\pi(f_j) - \pi(R_i^j)}{\pi(L_j^i) - \pi(R_i^j)}$$

3.4.2. *Evaluation over data informative range.* It can be argued that the performance measure should be entirely independent of model fit, and that evaluation should only be done for the metric value range for which we have information in the data. Having obtained $(F_j^{\min}, F_j^{\max})$ as previously, we could use the following measures instead of the previously suggested imputation approach,

1. For the unknown indicator corresponding to a right censored observation, we use the cumulative error measure,

$$ce = \frac{1}{R_j^i - F_j^{\min}} \int_{F_j^{\min}}^{R_i^j} e(\hat{f_j}) d(f_j).$$

2. For the unknown indicator corresponding to a left censored observation, we use the cumulative error measure,

$$ce = \frac{1}{F_j^{\max} - L_j^i} \int_{L_j^i}^{F_j^{\max}} e(\hat{f_j})d(f_j).$$

3. For the unknown indicator corresponding to an interval censored observation, we use the cumulative error measure,

$$ce = \frac{1}{F_j^{\max} - L_i^j + R_i^j - F_j^{\min}} \left\{ \int_{F_j^{\min}}^{R_i^j} e(\hat{f_j})d(f_j) + \int_{L_j^i}^{F_j^{\max}} e(\hat{f_j})d(f_j) \right\}.$$

We use both these proposed measures in our simulated and real data experiments for comparative evaluation.

3.5. *Proper scoring rules and other properties.* We discuss a few properties of the proposed measures. First of all, it is trivial to derive that in each case, the range of proposed measure is, $0 \leq ce \leq 1$.

Secondly, our proposed error rate estimator is a proper scoring rule. A scoring rule S is defined to be proper if an optimal strategy for the experimenter is to quote a distribution that matches their actual uncertainty. It is well known that the squared error loss and Brier score lead to proper scores, while the AuROCs in general, do not (Gneiting and Raftery, 2007). In our estimator, we show that the proper scoring is preserved, that is, if for any pair of biomechanical metrics, F^j, F^{j_1}, and observed data Y, we have,

$$\mathbb{E}_{F^j}[ce(Y, F^j)] \leq \mathbb{E}_{F^{j_1}}[ce(Y, F^j)$$

This implies that all other things remaining the same, our proposed method would always lead to the correct choice of metric under correct model specification.

Thirdly, our proposed estimator is asymptotically consistent, as given by the following theorem.

THEOREM 1. *Let $\pi(\cdot)$ be any uniformly consistent estimator for the lifetime distribution function $F(\cdot)$. Then the error estimate, $c(\hat{f_j})$ converges almost surely to $MSE(f_j)$ as $n \to \infty$. Additionally, if the metric in question F^j is uniformly bounded as a random variable, then all variants of the measure ce converge to the respective cumulative MSE with probability 1.*

4. Simulation Examples. We consider three simulation scenarios to assess the proposed measures in terms of ability to distinguish between candidate "time" metrics, for construction of injury probability curves. In the first scenario, we consider a set of 10 metrics, all of which are closely associated with injury status, but where the metrics are generated from distributions with differing variability. We compute the full range cumulative error estimators, the data informative range error estimators (dubbed fce and dce henceforth respectively), alongwith AuROC type measures. In the second scenario, we simulate 10 metrics, which are loosely associated with the injury status, from distributions with differing variability. In the final set, we simulate 10 more metrics, which have no association with injury status, and are randomly simulated from distributions with differing variability. For each of these cases, we simulate, 10, 20, 30 and 40 samples respectively. All the scores perform better with increasing sample sizes, though, clearly, both Brier score Type I and Brier score Type II perform substantially better than AuROC. Brier Score Type II seems to have a slight edge over Brier Score Type I. Further details of the simulation set-up is provided in the supplementary materials.

Sample Size	AUC	BS_Type_1	BS_Type_2
10	31%	64%	68%
20	33%	68%	67%
30	38%	72%	76%
40	54%	84%	89%

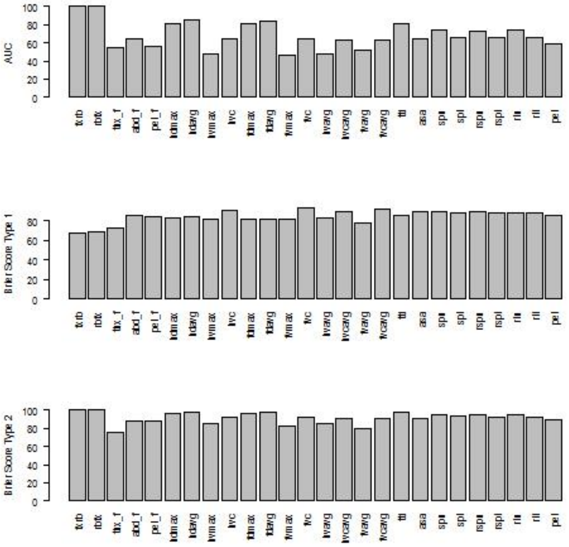
Table 1

Performance of different metrics. The percentages indicate number of times the best metric is chosen across simulations.)

5. Analysis of a PMHS experiment dataset. A published biomechanical dataset (Table 1) in Kuppa et al. (2003) is used to assess the each of performance evaluation methods along with AuROC. It consists of injury and noninjury information from side impact sled tests conducted at the Medical College of Wisconsin. They were obtained from whole body PMHS (Post Mortem Human Subject) tests conducted at different velocities, padding and rigid load wall conditions, offsets, and supplemental restraint systems, i.e., with and without side impact airbags. Specimens were subjected to single lateral impact loading. A total of 24 metrics are chosen from Table 2 in Kuppa et al. (2003). Biomechanical metrics included data obtained from different types of sensors: accelerometers for the thoracic trauma index (TTI), peak pelvic acceleration and Average Spine Acceleration (ASA) metrics, and load cells for the thoracic and pelvic forces. Unilateral or bilateral rib fractures in isolation or in combination with solid organ trauma occurred in some tests due to impact with the load wall, representing compression-related injury mechanism. Presence and absence of injury is done based on the Maximum Abbreviated Injury Scale (MAIS) scale, with injury severities greater than MAIS 3 classified as having an injury.

For the purpose of an easier visual comparison, we transform each of the two types of Brier scores by subtracting them from 1 and then multiplying them by 100, so that each is a score out of 100, with a higher score indicating better performance. The relative performances of the metrics are given in table 1 and also plotted in figure 1. While the AUC chooses TTI as the best metric among those available, Brier Score Type 1 chooses fvc and Brier Score Type 2 chooses TTI. Both of these metrics are valid biomechanical parameters that could be used for an injury curve. Based on our simulations, Brier Score Type II may be given priority. However Brier Score Type I also plays an important role - it helps vet metrics for which the parametric survival model has not converged - which are not picked up by AUC or Brier Score Type II - as has happened for the first couple of metrics - namely, fxrb and rbfx.

Fig 1. *Performance of different metrics. (Note: All scores are out of 100. The two Brier type Scores have transformed, so that in each case, higher is better.)*



	AUC	BS_Type_1	BS_Type_2
1	100.00	67.84	100.00
2	100.00	68.03	100.00
3	54.81	72.42	76.15
4	64.99	84.99	87.80
5	56.57	84.28	87.82
6	81.83	83.10	97.09
7	84.72	83.47	97.55
8	47.69	81.90	84.85
9	64.35	90.90	92.02
10	80.90	82.08	96.80
11	84.03	82.08	97.32
12	46.64	81.20	82.50
13	63.77	92.45	92.50
14	47.11	82.30	85.19
15	62.85	88.84	90.49
16	52.55	77.90	79.90
17	62.85	91.71	90.49
18	80.56	85.12	98.22
19	64.29	88.86	91.36
20	73.86	89.82	94.40
21	65.21	87.83	93.42
22	72.35	89.79	94.73
23	66.01	87.46	92.77
24	74.21	87.42	94.87
25	66.37	88.37	91.88
26	59.03	84.81	88.91

TABLE 2

Performance of different metrics. (Note: All scores are out of 100. The two Brier type Scores have transformed, so that in each case, higher is better.)

6. Concluding remarks. We develop a new measure for comparing performance for general survival models with right, left and/or interval censored data, an advantage over over ROC type metrics, which ignore censoring information. We have shown applicability of the proposed measures in PMHS experiments, where multiple biomechanical metrics are routinely gathered and/or derived during post processing. The present statistical analysis assists in the post processing 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 anthropomorphic test devices, routinely used to improve human safety and crashworthiness in automotive and military environments. The present methodology can therefore, be effectively used for dummy design, evaluation, injury criteria and human safety.

Beyond the current context, the proposed methods and results established are applicable in general "current status data" settings. The results establish properties and connections with scoring rules and could serve as benchmarks for estimation procedures in these contexts.

Acknowledgements. This material is the result of work supported with resources and use of facilities at the Zablocki VA Medical Center, Milwaukee, Wisconsin and the department of Neurosurgery at the Medical College of Wisconsin. This research was conducted as part of the WIAMan Project under contract N00024-13-D-6400, U.S. Army Research, Development and Engineering Command. This paper was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001436. Any views expressed in this article are those of the authors and not necessarily representative of the funding organizations.

References.

Berzuini, C. and Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in medicine* **13** 823–838.

Blanche, P., Dartigues, J.-F. and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine* **32** 5381–5397.

Blincoe, L., Miller, T. R., Zaloshnja, E. and Lawrence, B. A. (2015). The economic and societal impact of motor vehicle crashes, 2010 (Revised May 2015). *NHTSA Docket* **DOT HS 812 013**.

Chen, C.-M., Wei, J. C.-C., Hsu, C.-M. and Lee, M.-Y. (2014). Regression analysis of multivariate current status data with dependent censoring: application to ankylosing spondylitis data. *Statistics in medicine* **33** 772–785.

Danelson, K. A., Kemper, A. R., Mason, M. J., Tegtmeyer, M., Swiatkowski, S. A., Bolte, J. H., Hardy, W. N. et al. (2015). Comparison of ATD to PMHS Response in the Under-Body Blast Environment. *Stapp car crash journal* **59** 445–520.

Duchesne, T. (1999). *Multiple time scales in survival analysis*. University of Waterloo.

Fay, M. P. and Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: the interval R package. *Journal of Statistical Software* **36**.

Gallenberger, K., Yoganandan, N. and Pintar, F. (2013). Biomechanics of foot/ankle trauma with variable energy impacts. *Annals of advances in automotive medicine* **57** 123.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* **81** 618–623.

Gerds, T. A., Cai, T. and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical journal* **50** 457–479.

Gerds, T. A. and Schumacher, M. (2007). Efron-Type Measures of Prediction Error for Survival Analysis. *Biometrics* **63** 1283–1287.

Gerds, T. A., Kattan, M. W., Schumacher, M. and Yu, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* **32** 2173–2184.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* **18** 2529–2545.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography* **21** 498–507.

Kent, R. W. and Funk, J. R. (2004). Data Censoring and Parametric Distribution Assignment in the Development of Injury Risk Functions From Biochemical Data Technical Report, SAE Technical Paper.

Kleinberger, M., Sun, E., Eppinger, R., Kuppa, S. and Saul, R. (1998). Development of improved injury criteria for the assessment of advanced automotive restraint systems. *NHTSA Docket* **1998-4405** 9.

Kom, E. L., Graubard, B. I. and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American journal of epidemiology* **145** 72–80.

Kuppa, S., Eppinger, R. H., McKoy, F., Nguyen, T. et al. (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** 189.

Lobo, J. M., Jiménez-Valverde, A. and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography* **17** 145–151.

Ma, L., Hu, T. and Sun, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* asv020.

Mol, B., Coppus, S., Van der Veen, F. and Bossuyt, P. (2005). Evaluating predictors for the outcome of assisted reproductive technology: ROC-curves are misleading; calibration is not! *Fertility and Sterility* **84** S253–S254.

Nahum, A. M. and Melvin, J. W. (2012). *Accidental injury: biomechanics and prevention*. Springer Science & Business Media.

Oakes, D. (1995). Multiple time scales in survival analysis. *Lifetime Data Analysis* **1** 7–18.

Petitjean, A., Trosseille, X., Petit, P., Irwin, A., Hassan, J. and Praxl, N. (2009). Injury risk curves for the WorldSID 50th male dummy. *Stapp car crash journal* **53** 443.

Philippens, M., Wismans, J., Forbes, P., Yoganandan, N., Pintar, F. A. and Soltis, S. (2009). ES2 neck injury assessment reference values for lateral loading in side facing seats. *Stapp car crash journal* **53** 421.

Robbins, D., Melvin, J. and Stalnaker, R. (1976). The prediction of thoracic impact injuries Technical Report, SAE Technical Paper.

Sun, J. (1996). A non-parametric test for interval-censored failure time data with application to AIDS studies. *Statistics in medicine* **15** 1387–1395.

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)* 290–295.

Zhao, S., Hu, T., Ma, L., Wang, P. and Sun, J. (2015). Regression analysis of informative current status data with the additive hazards model. *Lifetime data analysis* **21** 241–258.