JOURNAL OF NEUROTRAUMA 38:435–445 (February 15, 2021)
Mary Ann Liebert, Inc.
DOI: 10.1089/neu.2020.7089

Accelerated Failure Time Survival Model to Analyze Morris Water Maze Latency Data

Clark R. Andersen,[1,2] Jordan Wolf,[1,3] Kristofer Jennings,[2] Donald S. Prough,[1,3] and Bridget E. Hawkins[1,3,4]

Abstract

Traumatic brain injury (TBI) induces cognitive deficits clinically and in animal models. Learning and memory testing is critical when evaluating potential therapeutic strategies and treatments to manage the effects of TBI. We evaluated three data analysis methods for the Morris water maze (MWM), a learning and memory assessment widely used in the neurotrauma field, to determine which statistical tool is optimal for MWM data. Hidden platform spatial MWM data aggregated from three separate experiments from the same laboratory were analyzed using 1) a logistic regression model, 2) an analysis of variance (ANOVA) model, and 3) an accelerated failure time (AFT) time-to-event model. The logistic regression model showed no significant evidence of differences between treatments among any swims over all days of the study, $p > 0.11$. Although the ANOVA model found significant evidence of differences between sham and TBI groups on three out of four swims on the third day, results are potentially biased due to the failure of this model to account for censoring. The time-to-event AFT model showed significant differences between sham and TBI over all swims on the third day, $p < 0.045$, taking censoring into account. We suggest AFT models should be the preferred analytical methodology for latency to platform associated with MWM studies.

Keywords: latency; learning and memory; Morris water maze; survival analysis; traumatic brain injury

Introduction

The Morris water maze (MWM) is a commonly used test for the assessment of learning and memory[1] for neurotrauma studies in rodents.[2,3] Its widespread use has caused the birth of multiple variations of Hamm's testing paradigms, resulting in a multitude of tests for learning and memory based on individual recipes of days of testing, visibility of platform, platform locations, entry sites, number of swims per day, intertrial intervals, cutoff swim times (maximum duration animal is allowed for completion of task), and outcome measures.

Most pre-clinical traumatic brain injury (TBI) animal studies utilizing the MWM have relatively small sample sizes,[4] typically fewer than 20 animals per group, and thus have relatively low power and may be easily influenced by outlier behavior of a few animals. Here, we have pooled the data from three prior studies at the University of Texas Medical Branch at Galveston (UTMB), focusing specifically on two groups, untreated sham versus fluid percussion TBI, to yield combined counts of 46 sham and 61 TBI animals. This provided us an opportunity to characterize the behavioral consequences of our injury model on MWM latency with unusually high power. Combining data sets to achieve increased power for analyses is discussed in greater detail in a publication by Hawkins and colleagues.[5]

In MWM studies, swim time (latency) to reach the submerged platform (the event of interest) within a fixed maximum duration (usually 120 sec) is fundamentally a time-to-event problem. Treating swim time alone as the outcome is erroneous because the swim time of those animals that fail to reach the platform within the available time is censored at the maximum duration, and statistical analyses that ignore the censoring (such as *t* tests or analysis of variance [ANOVA]) will yield potentially biased results; typically censored data in these circumstances is truncated at the maximum swim duration prior to analysis of either truncated data or averaged truncated data, which has the effect of yielding skewed distributions that yield downwardly biased estimates of the mean and variance, and these translate to corresponding downward biases in treatment effect sizes and invalid standard errors and *p*-values. The downward bias in effect size may be considered more conservative in a hypothesis testing context, but this is in conflict with the more liberal downward bias in standard errors and *p*-values, and it is not clear which bias may dominate results.

[1]The Moody Project for Translational Traumatic Brain Injury Research, [2]Office of Biostatistics, Department of Preventive Medicine and Population Health, [3]Department of Anesthesiology, [4]School of Nursing, University of Texas Medical Branch, Galveston, Texas, USA.

An alternative is to utilize the binary state of whether or not the animal found the platform as an outcome; assuming relative homogeneity of the conduct of the trial, analyses of this binary outcome (such as by chi-square test or logistic regression) are unbiased but have unacceptably low power. Time-to-event data are appropriately modeled by utilizing both the time and event information simultaneously. This approach is commonly referred to as ''survival modeling'' although in the context of MWM the event of interest is reaching the submerged platform, rather than mortality. In this article, we compare and contrast these three approaches with the goal of reinforcing the point that time-to-event models are the most appropriate analytical method for MWM latency.

Time-to-event models of interest include the non-parametric Kaplan-Meier method[6] paired with the log-rank test, which is analogous to a two-sample *t* test of continuous data. This approach is valid when analyzing data from a single swim trial on a single day but is inadequate when the same animals have multiple swims (repeated measures) or with additional variables of interest such as the day of swim and swim per day. The popular semiparametric Cox proportional hazards model can model time-to-event with relation to multiple variables and cope with clustering.[7] Similarly, we consider the parametric accelerated failure time (AFT) model as an alternative that does not require the proportional hazards assumption. The Cox model requires that the proportional hazards assumption is valid. Although the AFT model does not require the proportional hazards assumption, it does require selection of the correct parametric distribution from among a pool of common alternatives. With these modeling approaches, as with linear models, differences among the groups and time-points may be estimated by contrasts, typically adjusted for multiple comparisons, and summarized in tables. For the Cox model, taking the exponential of the reported coefficient for the contrast estimate yields the hazard ratio between the groups. For the AFT model, using the corresponding exponentiated coefficient provides the ratio of the (model-adjusted) mean swim times between the groups.

Sometimes animals are unable to swim and are removed from the MWM tank to avoid drowning. In the time-to-event context, such animals are censored at the time they are removed from the pool. They do not experience the ''event'' of reaching the platform, but their experience of swimming up until removal can still contribute information to the model. In other approaches that are unable to account for censoring (for instance, logistic regression for incidence of reaching the platform or ANOVA for swim duration), investigators may choose to exclude this animal's swim from the analysis or treat it as a successful swim to the platform, but either alternative introduces bias.

The utilization of time-to-event models to compensate for the bias due to censoring in studies of MWM latency has been considered[8,9] and has been applied in other studies.[10,11] Those studies handled time-to-event modeling with Cox proportional hazards models. Here, we argue that the proportional hazards assumption may not be satisfied in the MWM context, particularly for our data, when modeling comparisons both between swims and between trial days. We suggest the use of AFT models instead. AFT models fitted to appropriate distributions have properties similar to Cox models, but without the assumption of proportional hazards, producing estimates of model-adjusted mean swim durations rather than hazards. We argue that the ratio between mean swim times facilitates a more intuitive interpretation than the ratio between hazards. Additional functionality has been developed for Cox models, beyond those currently available for AFT models, including allowance for left censoring, interval censoring, and repetitive events per trial, but these capabilities are not required in the analysis of MWM latency.

This discussion of analysis of latency is focused on studies of latency in the context of swimming to find a hidden platform; latency in the absence of a hidden platform (probe trials) is a separate topic not addressed in this article, which is adequately modeled by conventional ANOVA.

Methods

Animals

These studies were conducted in a facility approved by the American Association for the Accreditation of Laboratory Animal Care (AAALAC). All experiments were performed in accordance with the National Institutes of Health *Guide for the Care and Use of Laboratory Animals* (8th edition, National Research Council) and approved by the Institutional Animal Care and Use Committee (IACUC) of UTMB. Adult, male, Sprague-Dawley rats (Charles Rivers Laboratories, Inc., Portland, ME, USA), 250–400 g, were group housed (two rats of similar injury status per cage) and had access to food and water *ad libitum* in a vivarium with these constant conditions: light cycle (0600–1800), temperature (21°C–23°C), and humidity (40%–50%). Unless noted, all animals were provided with enrichment materials, such as a cardboard tube, in their home cage.

Fluid percussion injury

All animal surgeries were performed under aseptic conditions by trained investigators in accordance with our IACUC-approved protocol, minimizing pain and distress at all times. Animals were anesthetized with 4% isoflurane in an anesthetic chamber, intubated, and mechanically ventilated with 1.5–2.0% isoflurane in O_2: room air (70:30) using a volume ventilator (EDCO Scientific, Chapel Hill, NC, USA). Rats were prepared for parasagittal fluid-percussion TBI (FPI) as previously described.[12] Briefly, animals were placed in a stereotaxic head frame and the scalp was sagittally incised. A 4.0-mm diameter hole was trephined into the skull 2.0 mm to the right of the sagittal suture and midway between lambda and bregma, and then a modified 20-gauge Luerlok syringe hub (Becton-Dickinson, Franklin Lakes, NJ, USA) was placed over the exposed dura, bonded in place with cyanoacrylic adhesive and covered with dental acrylic. Animals with punctured dura were excluded from the study. Prior to FPI induction, the device and connector were filled with sterile degassed water and checked for air bubbles. The device was prepared for the injury by delivering approximately three test pulses (confirmed by a smooth waveform on the oscilloscope) while the Luerlok at the end of the tubing was in the closed position.

Just prior to FPI induction, isoflurane was temporarily discontinued and rats were connected to the fluid percussion trauma device (Custom Design and Fabrication, Virginia Commonwealth University, VA, USA). They were subjected to FPI (266–320 mV oscilloscope [Tektronix TDS 1002 60 MHz, two-channel digital real time with Trauma Inducer Pressure Transducer Amplifier] readings, 1.81–2.17 atm range calculated, consistently held at 15.5-cm pendulum height, and pressure pulse length set at 25 msec) immediately after the return of a withdrawal reflex to paw pinch. The withdrawal reflex is a spinal reflex that returns after the cessation of anesthetics prior to the return of higher-level reflexes (e.g., righting reflex [RR]) while the rat remains unconscious. Because FPI is administered immediately after a withdrawal reflex is detected, rats are unconscious at the time of injury. It is necessary to discontinue the isoflurane immediately prior to injury to reduce the effects of anesthesia on the time required for the rat to right itself from a supine position (RR). The RR is a brainstem reflex that returns prior to thalamocortical function during recovery from unconsciousness due to anesthesia or brain injury.

After FPI or sham injury, rats were disconnected from the fluid percussion device and RR was assessed until a normal RR was observed three times (and the time at third righting was recorded). Rats were then placed on 2% isoflurane while wound sites were infused with bupivicaine and skin was closed with wound clips. The animals received approximately 100 mg/kg acetaminophen suppository before emerging from anesthesia. Isoflurane was discontinued and the rats were extubated and allowed to recover in a warm, humidified incubator. When each rat was fully recovered, it was returned to its home cage with *ad libitum* food and water. Each rat (two per cage) was housed with a cagemate of similar injury status (naives were housed with other naives, shams were housed with other sham-injured rats, and FPI-injured rats were housed with other FPI-injured rats) to prevent anxiety.

All animals were monitored for signs of infection, severe neurological injury, or discomfort. Signs of discomfort or pain in rodents include persistent dormouse position and unwillingness to move, refusal to eat or drink, vocalizations when handled, posturing, aggressiveness, and polyphagia of bedding. Rats exhibiting these symptoms were humanely euthanized immediately (4% isoflurane in an anesthetic chamber followed by decapitation) to prevent pain and distress. Any rats that received a return of RR time of less than 20 min or animals that experienced neurogenic pulmonary edema immediately after the FPI were excluded from the study. All sham control animals received the same amount of anesthesia and were prepared identically to the injured animals with the exception of the actual FPI injury.

Behavior studies

Immediately following FPI in rodents, the RR (the reflex for the animal to turn over to its normal upright position when it is placed on its back) is suppressed and its return is known to be a clinical correlate for return to consciousness in patients who have sustained more than a mild TBI.[15] The RR time was recorded when the rat had righted itself three times consecutively after being placed on its back. In this study, the RR times for all of the injured animals fell between 22 and 26 min, indicating greater than mild impairment and in comparison, the RR for the sham-injured controls (receiving identical surgical preparation and anesthesia levels as the FPI-injured rats) was under 10 min in length.

Neuroscore evaluation

Animals received neuroscore (NS) evaluations pre- and post-FPI. Behavioral competency of each animal was assessed prior to surgery by testing its baseline reflex performance with the NS evaluation, as described previously.[14,15] Animals were excluded from study if any deficits were present before surgery.

Morris water maze

The MWM is a commonly used test for the assessment of learning and memory for neurotrauma studies in rodents.[2] The testing paradigm and equipment used was similar to that reported in a study by Nichols and associates,[16] with the exception that testing was for 3 days instead of 5 days (post-injury days 11–13) and the probe trial (platform was removed from the pool and memory was tested using a 30-sec swim to find time spent in target quadrant as a correlate of memory retention and recall in the absence of the escape platform) was post-injury day 15 (probe data not shown). The MWM tank (2 ft. height and 70 in. diameter) was located in a quiet room with a camera mounted on the ceiling and a computer with an ANYMaze tracking system located on the other side of a curtain where the experimenter was while the rat was swimming. The room had a bookshelf and items mounted on each wall that remained consistently in place throughout the experiments and the water maze was drained and cleaned at the end of each day (in between each swim, feces were removed and water was swirled to prevent scent trails to the platform). During the experiment, the overhead lighting was turned off and a floodlight that remained in the same corner of the room was turned on. In between swim trials, the rat would recover in a warming chamber for 4 min. The water temperature was maintained at 26°C (± 1 degree in either direction).

Each rat was placed in the water facing the tank from one of the four equally spaced entry points along the perimeter of the tank. Once the rat was released, the ANYMaze system tracked the movement of the rat and stopped the trial once the animal reached the hidden platform (clear plexiglass; 4.5 in. diameter) or once the maximum swim trial time was reached. To prevent exhaustion from multiple swims per day, we set the maximum time for each swim at 120 sec. The rats that were unable to find the platform were led to it and placed on it for 30 sec before being carried to the heated recovery chamber. All rats swam four swim trials (one from each entry point) per day for 3 days (platform remained in the same location for every trial). Behavior testing was performed by the same technician for all time-points in the study. The entry points were randomly selected (for instance, the first rat's first swim could result in one of four possible entry-point options [triangle, square, circle, or X], but for that rat's second trial, only one of three entry points were possible, and so forth, such that each rat started from all entry points each day) for each animal's trial and the behavior technician was blinded to injury status and intervention.

Statistical analyses

This study combines sham and TBI data from three prior experiments, each with similar experimental design and a common injury model. Experiment A contributed data from 8 animals per group (sham vs. TBI); experiment B contributed data from 13 animals per group; and experiment C contributed data from 25 sham and 40 TBI animals. Altogether the total cohort includes 46 sham and 61 TBI animals, with latency measures for four swims per day over 3 sequential days [days 11–13 post-injury].

Univariate summaries by group, day, and swim per day include the mean and standard deviation of swim duration, as well as counts and percentages of animals that successfully reached the submerged platform.

A time-to-event model was used to relate the swim duration (latency) to find the submerged platform (event) to swim day (days 1, 2, 3), swim per day (swims 1, 2, 3, 4), treatment (sham vs.TBI), and interactions among these variables, while adjusting for experiment (experiments A, B, C) and clustering (similar to that in mixed effect models) on animal to control for repeated measures of the same animals over multiple swims.

A semiparametric Cox proportional hazards model was initially considered as the time-to-event model[17]; however, the proportional hazards assumption was violated both in plots of hazard curves over time[18] and in formal tests based upon Schoenfeld residuals.[19] Nonproportionality was observed between days, between swims within the same day, and between treatment groups.

Considering the violations of the proportionality assumption, we opted instead to use an AFT model.[20] AFT models were fit using Weibull, exponential, Gaussian, logistic, lognormal, and log-logistic distributions. The model with log-logistic distribution had the lowest Akaike information criterion (AIC) and was selected as optimal for our analysis. Differences between sham and TBI treatments by swim per day for each day were estimated by Hommel-adjusted contrasts.[21] Differences between days by group and swim were similarly estimated.

Additionally, to allow comparison of the time-to-event model with more common methodologies, a mixed effect ANOVA was used to model swim duration (ignoring censoring bias, such that swims exceeding 120 sec were set as 120 sec), and a mixed effect logistic regression modeled incidence of reaching the submerged

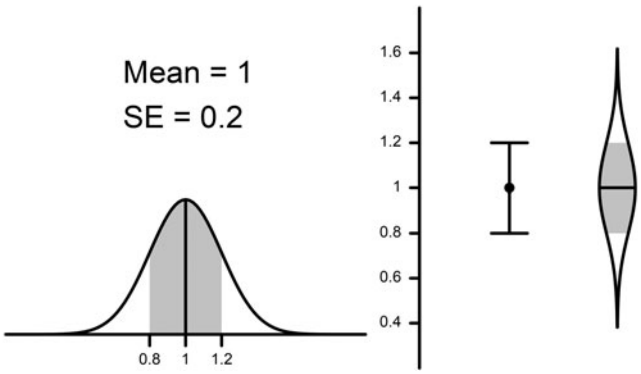

FIG. 1. Interpretation of catseye plots (see article by Cumming[26]). Catseye plots (so-named due to their visual similarity to a cat's eye) are used to illustrate the estimated distribution of the model-adjusted mean, which is normally distributed with a standard deviation equal to the standard error (SE). This example illustrates the representation of a mean of 1 with an SE of 0.2, shown as a conventional normal distribution ''bell'' curve at left with ±SE shaded, and at right as a conventional point with ±SE confidence interval, followed by the corresponding catseye plot with shaded ±SE interval. The outlined area of the catseye encapsulates 99.8% of the normal distribution of the mean (compare it with the normal curve at left, and envision that curve rotated 90 degrees and reflected about its axis), with the intention being to provide a more complete sense of the distribution of the estimate of the mean than would be feasible with just a ±SE or 95% interval.

platform. As with the time-to-event model, each of these models were with relation to swim day, swim per day, treatment, and interactions among these variables, while adjusting for experiment and blocking on animal to control for repeated measures. Likewise, differences between sham and TBI treatments by swim per day for each day were estimated by Hommel-adjusted contrasts.

Statistical analyses were performed using R statistical software.[22] In all statistical tests, alpha = 0.05. The ''survival'' package was used for both Cox and AFT time-to-event modeling.[17,23] Hazard functions were estimated using the ''muhaz'' package.[24] Differences among factor levels in the models were estimated using the ''emmeans'' package.[25] Catseye plots[26] were produced using the ''catseyes'' package[27] and are explained in Figure 1. R code and data sets used are included in the supplementary materials: Supplementary Modeling S1 and S2, and Supplementary Data Sets S1, S2, S3, and S4.

Results

Univariate summaries of swim duration (latency) and incidence of reaching the hidden platform by treatment group, day, and swim per day are provided in Table 1. Maximum swim duration was truncated at 120 sec, and associated summary statistics reflect this censoring bias. The table shows the expected trend in reduced latency and increased success rate of reaching the platform over time, as well as relatively increased latency and reduced success rate for the TBI group in comparison with sham. Figure 2 illustrates the same trends graphically on a per-rat basis for each day and swim and highlights the censoring at 120 sec.

The time-to-event AFT model showed significant differences between treatments over all swims on the third day, $p < 0.045$, as summarized in Table 2 and Figure 3. The mean swim duration of

Table 1. Univariate Summaries of Swim Duration and Incidence of Reaching Platform by Group, Day, and Swim per Day

				Swim duration							*Reached platform*	
Group	*Day*	*Swim*	N	*Mean*	*SD*	*Median*	*Q1*	*Q3*	*Min.*	*Max.*	N	%
Sham	1	1	46	102.7	34.8	120.0	108.5	120.0	4.9	120.0	13	28
Sham	1	2	46	84.3	42.0	116.8	47.7	120.0	4.0	120.0	25	54
Sham	1	3	46	69.5	43.9	69.6	28.9	120.0	3.0	120.0	30	65
Sham	1	4	46	51.2	43.1	32.1	15.1	97.9	6.2	120.0	37	80
Sham	2	1	46	61.1	45.1	53.3	19.2	119.9	3.2	120.0	34	74
Sham	2	2	46	51.9	40.5	45.4	18.4	85.0	1.3	120.0	38	83
Sham	2	3	46	34.2	33.1	20.6	9.5	46.8	3.6	120.0	43	93
Sham	2	4	46	29.6	32.1	17.8	6.2	46.0	1.3	120.0	44	96
Sham	3	1	46	35.2	35.4	19.7	9.5	45.9	4.7	120.0	42	91
Sham	3	2	46	23.8	24.7	13.4	5.7	33.2	0.7	120.0	45	98
Sham	3	3	46	19.7	13.6	14.9	10.7	25.8	3.1	60.3	46	100
Sham	3	4	46	18.4	20.9	12.0	6.7	18.9	1.3	117.7	46	100
TBI	1	1	61	105.0	34.7	120.0	120.0	120.0	2.2	120.0	11	18
TBI	1	2	61	91.2	38.8	120.0	67.1	120.0	4.5	120.0	29	48
TBI	1	3	61	82.1	39.4	86.1	55.7	120.0	1.6	120.0	38	62
TBI	1	4	61	64.8	43.9	55.5	22.6	120.0	4.9	120.0	42	69
TBI	2	1	61	69.6	44.6	63.7	27.3	120.0	1.3	120.0	40	66
TBI	2	2	60	61.8	44.8	48.3	24.1	120.0	3.0	120.0	43	72
TBI	2	3	60	51.5	38.8	41.3	19.0	76.3	5.2	120.0	52	87
TBI	2	4	60	49.1	43.8	29.3	13.4	95.8	1.3	120.0	47	78
TBI	3	1	61	57.6	43.2	43.8	17.7	120.0	2.8	120.0	45	74
TBI	3	2	61	51.0	38.6	37.2	21.4	76.5	0.6	120.0	53	87
TBI	3	3	61	46.5	38.6	32.4	12.9	76.8	3.3	120.0	53	87
TBI	3	4	61	32.0	31.5	22.8	10.3	45.3	1.3	120.0	58	95

SD, standard deviation; TBI, traumatic brain injury.

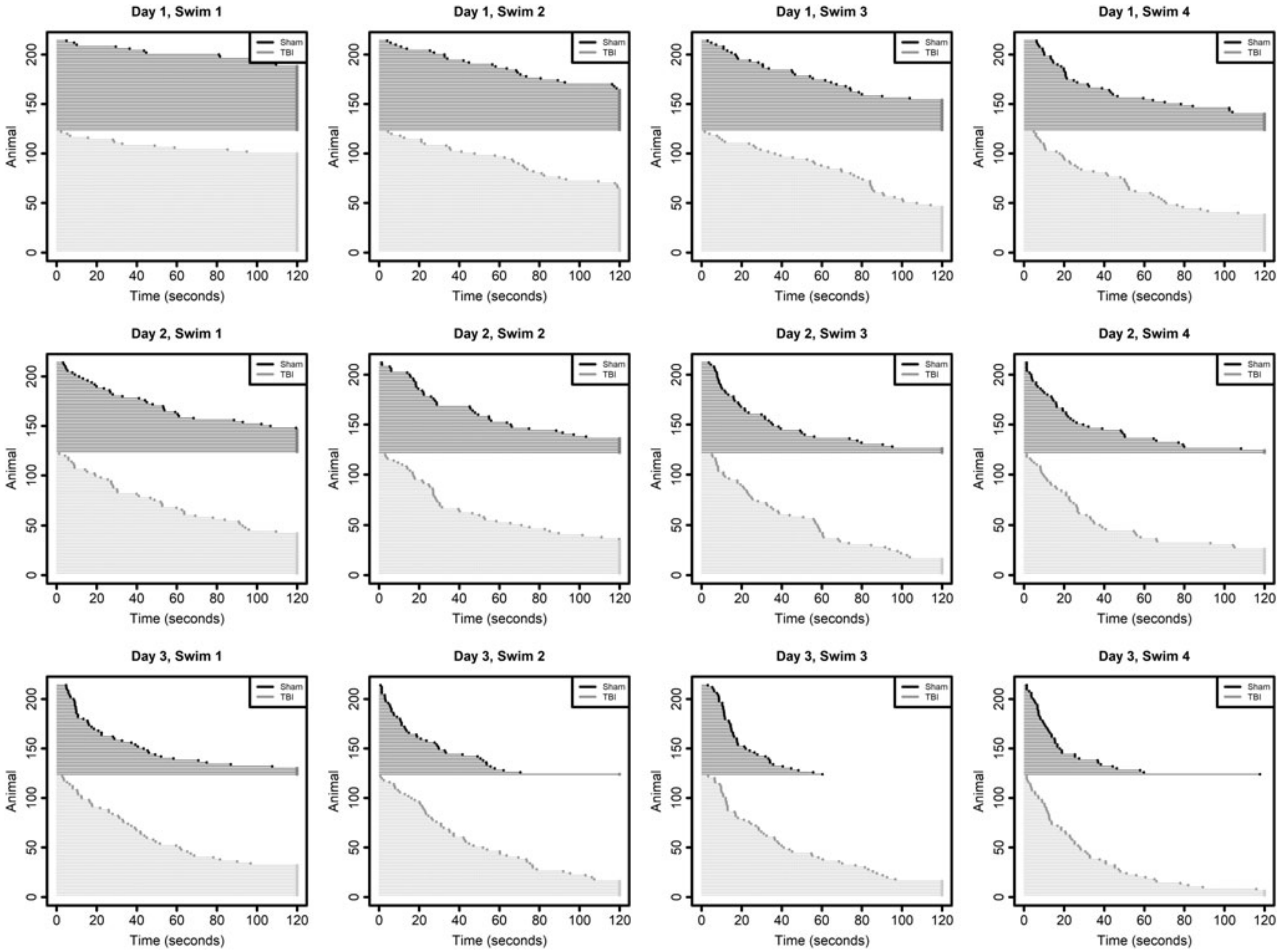

FIG. 2. Swim duration of each animal, for each day and swim per day, sorted by treatment group and duration. Each horizontal line tracks the swimming time of a single animal from 0 sec to completion of the swim. Successfully reaching the platform is indicated by a closed circle if prior to 120 sec, or an open circle if censored at 120 sec (which may be difficult to see due to the large number of animals shown). These plots are loosely analogous to Kaplan-Meier plots in the absence of explicit representation of probability, and with all censoring occurring at 120 sec.

TABLE 2. TIME-TO-EVENT (REACHING PLATFORM) MODEL-ADJUSTED DIFFERENCES BETWEEN TBI AND SHAM, BY DAY AND SWIM PER DAY

Day	*Swim*	*Estimate*	*SE*	*Exp (estimate)*	*CI 95 min.*	*CI 95 max.*	P-*value*	*Hommel* p-*value*
1	1	0.33	0.37	1.39	0.68	2.85	0.3702	0.3969
1	2	0.22	0.26	1.25	0.75	2.10	0.3969	0.3969
1	3	0.36	0.26	1.43	0.86	2.38	0.1666	0.3969
1	4	0.46	0.26	1.59	0.95	2.65	0.0753	0.3969
2	1	0.29	0.29	1.34	0.76	2.35	0.3163	0.3969
2	2	0.26	0.25	1.29	0.79	2.12	0.3060	0.3969
2	3	0.58	0.22	1.78	1.16	2.73	**0.0080**	0.0559
2	4	0.73	0.28	2.07	1.21	3.56	**0.0084**	0.0587
3	1	0.75	0.25	2.11	1.31	3.42	**0.0023**	**0.0211**
3	2	1.04	0.25	2.82	1.75	4.57	**0.0000**	**0.0003**
3	3	0.75	0.19	2.12	1.46	3.08	**0.0001**	**0.0009**
3	4	0.56	0.21	1.76	1.17	2.64	**0.0068**	**0.0474**

Exp (estimate) provides the ratio of model-adjusted mean swim durations between TBI and Sham groups. The Hommel-adjusted *p*-values compensate for multiple testing. By the third swim of day 2, the TBI group had about double the latency of the Sham group, although this doesn't show significance until day 3. Bold = $p < 0.05$.

CI, confidence interval; SE, standard error; TBI, traumatic brain injury.

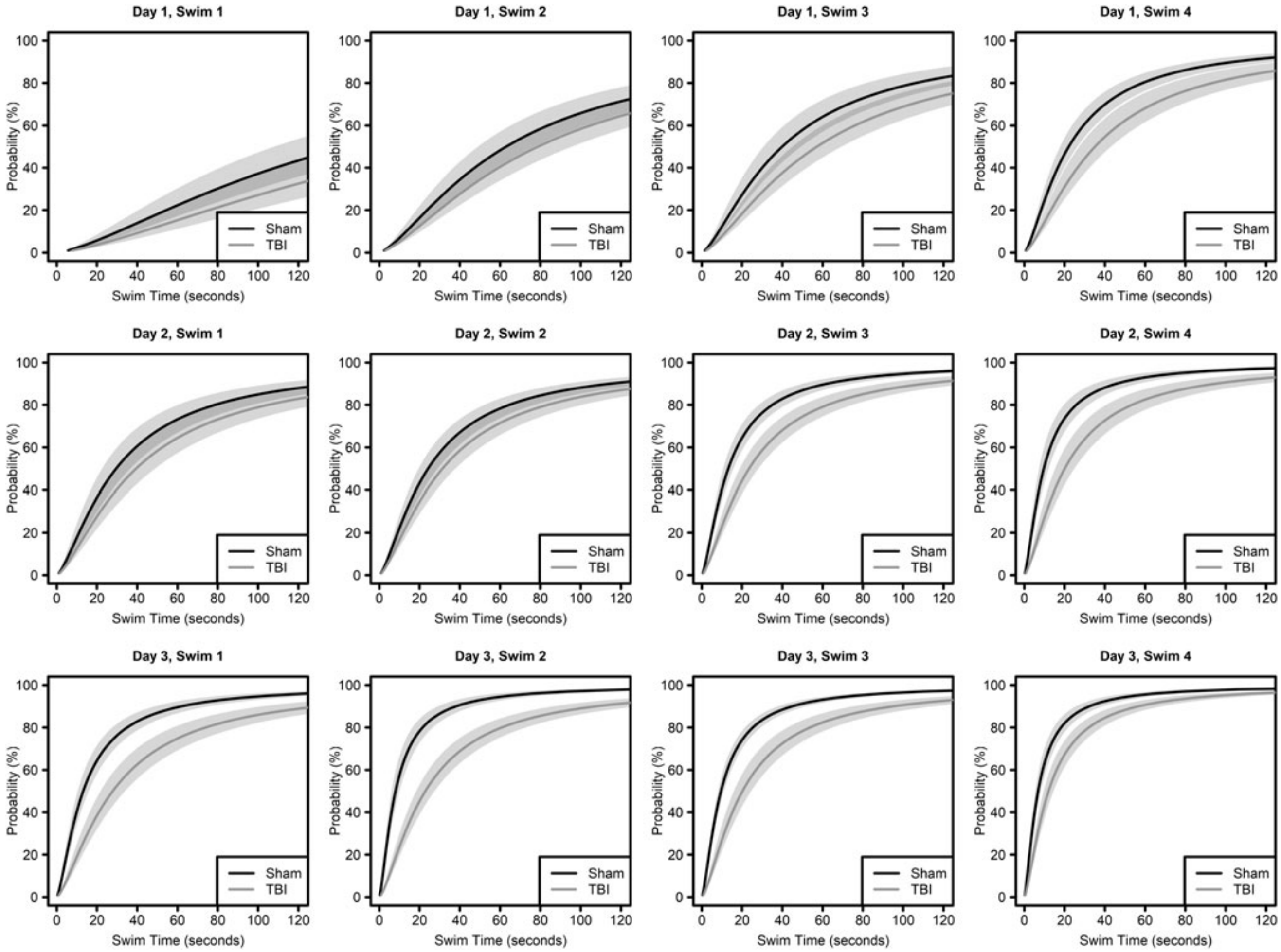

FIG. 3. Probability of reaching the hidden platform over time for each treatment group, per predictions from the accelerated failure time (AFT) time-to-event model, by day and swim per day. Shaded intervals indicate ±SE (standard error) over time (horizontal intervals, rather than vertical).

the TBI group was approximately double that of the sham throughout day 3, and about the same during the latter swims of day 2, although the difference was not significant. The figure shows a consistent trend of lower probability of reaching the platform over time for the TBI group in comparison with the sham group. It also shows the expected trend in increased probability of finding the platform over time both over multiple swims per day and between days. Table 3 shows broadly significant trend reductions in mean swim time over subsequent days for each group and swim per day, generally with greater reductions for the sham group than TBI.

The AFT model also found significant evidence of differences associated with the experiments that were the source of the data used in this analysis (table excluded for brevity, because adjustment for the source experiment was incidental to the analysis). Mean latency from source experiments B and C were approximately double that in experiment A (2.1 and 1.7 times, respectively, with Hommel-adjusted $p = 0.0001$ and $p = 0.0012$); there was no evidence of any difference between experiments B and C.

The ANOVA model found significant evidence of differences between treatments over the first 3 swims of day 3, $p < 0.03$, although not on the fourth swim of that day, as summarized in Table 4 and Figure 4. The TBI group averaged about 25 sec greater latency than the sham group over these swims. These ANOVA results should be considered cautiously, as they are potentially biased due to the failure of this model to account for censoring.

The logistic regression model showed no significant evidence of differences between treatments among any swims over all days of the study, $p > 0.11$, as summarized in Table 5 and Figure 5. The logistic model estimates of differences among treatments were problematic for swims 3 and 4 of day 3 due to 100% of the animals in the sham group reaching the platform on those swims. As a result, those estimates are unreliable.

Discussion

The time-to-event AFT model, as summarized in Tables 2 and 3 and illustrated in Figure 3, tells a story of impairment due to TBI, as well as the process of learning over the course of each day, and from day to day. We find it helpful to represent the results of the time-to-event model graphically by plotting the probability of reaching the platform with relation to swim time, separately by group, with a separate figure for each swim trial; anecdotally, clinicians find this representation intuitive, as they are accustomed to seeing it in the context of survival curves illustrating patient mortality.

The ANOVA model was problematic, not only because the model was biased by censoring, but also because the distribution of

Table 3. Time-to-Event (Reaching Platform) Model-Adjusted Differences between Days, by Treatment Group and Swim per Day

Contrast	*Group*	*Swim*	*Estimate*	*SE*	*Exp (estimate)*	*CI 95 min.*	*CI 95 max.*	P-*value*	*Hommel* p-*value*
2 - 1	Sham	1	-1.59	0.34	0.20	0.10	0.40	<0.0001	**<0.0001**
3 - 1	Sham	1	-2.39	0.33	0.09	0.05	0.18	<0.0001	**<0.0001**
3 - 2	Sham	1	-0.81	0.27	0.45	0.26	0.76	0.0029	**0.0236**
2 - 1	Sham	2	-0.96	0.25	0.38	0.24	0.62	0.0001	**0.0012**
3 - 1	Sham	2	-2.05	0.26	0.13	0.08	0.21	<0.0001	**<0.0001**
3 - 2	Sham	2	-1.10	0.26	0.33	0.20	0.55	<0.0001	**0.0002**
2 - 1	Sham	3	-1.11	0.24	0.33	0.21	0.53	<0.0001	**<0.0001**
3 - 1	Sham	3	-1.43	0.24	0.24	0.15	0.38	<0.0001	**<0.0001**
3 - 2	Sham	3	-0.32	0.19	0.73	0.50	1.05	0.0909	0.2124
2 - 1	Sham	4	-0.82	0.27	0.44	0.26	0.74	0.0022	**0.0176**
3 - 1	Sham	4	-1.17	0.23	0.31	0.20	0.49	0.0000	**<0.0001**
3 - 2	Sham	4	-0.34	0.22	0.71	0.46	1.08	0.1119	0.2238
2 - 1	TBI	1	-1.63	0.32	0.20	0.11	0.37	<0.0001	**<0.0001**
3 - 1	TBI	1	-1.98	0.30	0.14	0.08	0.25	<0.0001	**<0.0001**
3 - 2	TBI	1	-0.35	0.23	0.71	0.45	1.10	0.1227	0.2454
2 - 1	TBI	2	-0.92	0.24	0.40	0.25	0.64	0.0001	**0.0013**
3 - 1	TBI	2	-1.24	0.21	0.29	0.19	0.44	<0.0001	**<0.0001**
3 - 2	TBI	2	-0.32	0.22	0.73	0.48	1.11	0.1416	0.2832
2 - 1	TBI	3	-0.89	0.20	0.41	0.28	0.60	<0.0001	**0.0001**
3 - 1	TBI	3	-1.04	0.22	0.35	0.23	0.55	<0.0001	**0.0001**
3 - 2	TBI	3	-0.15	0.16	0.86	0.63	1.18	0.3460	0.3460
2 - 1	TBI	4	-0.56	0.24	0.57	0.36	0.92	0.0204	0.1222
3 - 1	TBI	4	-1.07	0.22	0.34	0.22	0.52	<0.0001	**<0.0001**
3 - 2	TBI	4	-0.51	0.20	0.60	0.40	0.90	0.0129	0.0774

Exp (estimate) provides the ratio of model-adjusted mean swim durations between TBI and Sham. The Hommel-adjusted *p*-values compensate for multiple testing. Bold = $p < 0.05$.

CI, confidence interval; SE, standard error; TBI, traumatic brain injury.

the durations was not particularly normal, as is visually clear per the scatterplots in Figure 4. The proportion of swims that are censored due to the threshold of 120 sec (maximum swim time per animal per swim) is highest on the first swim of each day, then declines with subsequent swims, as well as with subsequent days. The proportion censored also varies between groups due to treatment. Bear in mind that even non-parametric tests such as Mann-Whitney require that each group have the same distribution, and simple parametric methods such as *t* test and ANOVA assume normal distribution of the means and homogeneous variance, and this type of data satisfies none of these assumptions.

Table 4. ANOVA Model-Adjusted Differences in Latency between TBI and Sham, by Day and Swim Per Day

Day	*Swim*	*Estimate*	*SE*	P-*value*	*Hommel* p-*value*
1	1	2.30	7.41	0.7566	0.7566
1	2	6.85	7.41	0.3575	0.7150
1	3	12.56	7.41	0.0932	0.4255
1	4	13.55	7.41	0.0704	0.3522
2	1	8.48	7.41	0.2553	0.5363
2	2	9.92	7.44	0.1850	0.5363
2	3	17.37	7.44	**0.0214**	0.1714
2	4	19.48	7.44	**0.0101**	0.0912
3	1	22.39	7.41	**0.0032**	**0.0318**
3	2	27.06	7.41	**0.0004**	**0.0045**
3	3	26.74	7.41	**0.0005**	**0.0053**
3	4	13.53	7.41	0.0708	0.3539

Hommel *p*-values adjust for multiple testing. This model ignores censoring associated with truncation of swim times at the 120-sec maximum, so these estimates should be considered biased. Bold = $p < 0.05$.

ANOVA, analysis of variance; SE, standard error; TBI, traumatic brain injury.

Due to the small numbers of animals in most studies, even the convergence to normal distribution of the mean due to the central limit theorem cannot be assumed. Transformations such as a log transformation may have been helpful on particular days and swims per day, but no single transformation would simultaneously correct the skewness over all swims, and none would correct the censoring. The distributions are not only skewed, but also heterogeneous. These considerations suggest that a conventional ANOVA, treating duration as a continuous variable with homogeneous normal distribution, is fundamentally an incorrect model for MWM latency data. If applied to MWM latency data, the results will be biased and standard errors and *p*-values will be invalid.

The logistic regression model is a valid model for the incidence of animals finding the platform, as it incorporates censoring information and is not subject to the distributional problems associated with latency. Unfortunately, as made clear by comparing the estimates of treatment effect from the logistic model (Table 5) with those of the AFT or ANOVA models (Tables 2 and 4), it is not a particularly powerful model and has difficulty whenever incidence for any one group is near the extremes of 0% or 100%. This latter issue might be addressed by utilizing something like a Firth-penalized regression.[28]

The significant differences that the AFT model found among the source experiments of the data that were pooled for this analysis serve to reinforce the importance of randomization in individual experiments and the need to account for the evolving changes in

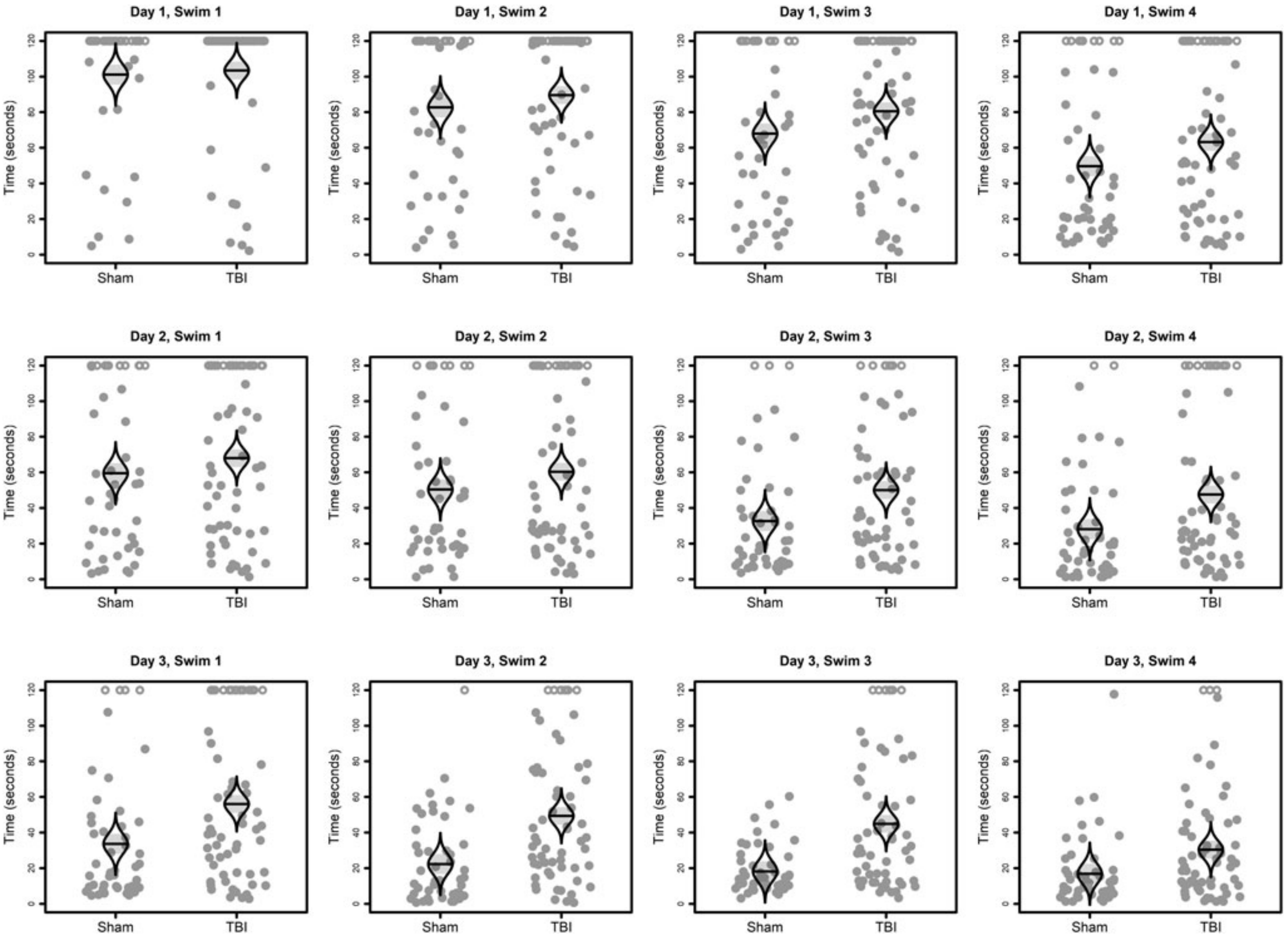

FIG. 4. Model-adjusted mean swim durations for each treatment group, per predictions from the analysis of variance (ANOVA) model, by day and swim per day. The distribution of the means is illustrated by catseye plots (see discussion in Fig. 1), with shaded ±SE (standard error). Single-animal swim durations are indicated by scatterplot circles (which have been randomly jittered horizontally), with a closed circle if the platform was reached prior to 120 sec, or an open circle if censored at 120 sec (may be difficult to see due to the large number of animals shown). Note that estimates from the ANOVA model are biased due to censoring.

Table 5. Logistic Regression Model-Adjusted Differences in Incidence of Reaching Platform between TBI and Sham, by Day and Swim per Day, with Hommel-Adjusted *P*-Values to Compensate for Multiple Testing

Day	*Swim*	*Estimate*	*SE*	*Odds ratio*	*CI 95 min.*	*CI 95 max.*	P-*value*	*Hommel* p-*value*
1	1	-0.67	0.48	0.51	0.20	1.30	0.1612	0.8836
1	2	-0.32	0.42	0.72	0.32	1.65	0.4418	0.9997
1	3	-0.14	0.43	0.87	0.37	2.03	0.7412	0.9997
1	4	-0.70	0.48	0.50	0.20	1.27	0.1494	0.8836
2	1	-0.45	0.45	0.64	0.26	1.55	0.3228	0.9997
2	2	-0.71	0.49	0.49	0.19	1.30	0.1568	0.8836
2	3	-0.79	0.68	0.46	0.12	1.71	0.2474	0.9997
2	4	-1.93	0.74	0.15	0.03	0.62	**0.0103**	0.1134
3	1	-1.44	0.58	0.24	0.08	0.75	**0.0157**	0.1723
3	2	-1.94	0.98	0.14	0.02	0.98	**0.0498**	0.4480
3	3	-25.45	65375.01	0.00	0.00	Inf	0.9997	0.9997
3	4	-24.23	65374.69	0.00	0.00	Inf	0.9997	0.9997

Exp (estimate) gives the odds ratio, which provides the odds of reaching the platform for TBI as opposed to Sham groups. Bold = $p < 0.05$.
CI, confidence interval; Inf, infinity; SE, standard error; TBI, traumatic brain injury.

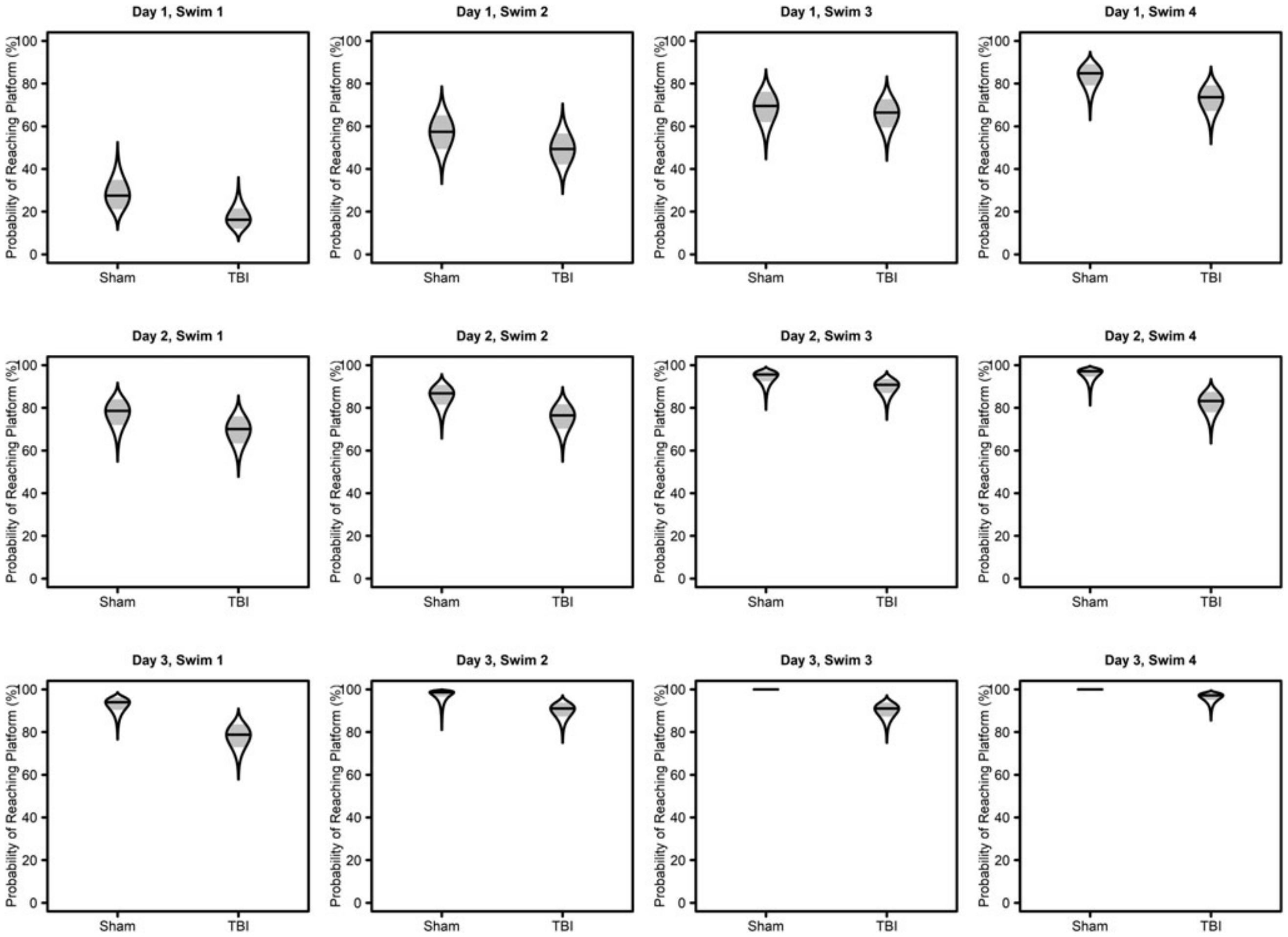

FIG. 5. Probability of reaching the hidden platform over time for each treatment group, per predictions from the logistic regression model, by day and swim per day. The distribution of the model-adjusted means is illustrated by catseye plots (see discussion Fig. 1), with shaded ±SE (standard error), and have been transformed from the logit scale to the probability scale such that distributions near 0% or 100% are asymmetrically distorted accordingly.

animal handling, handlers, laboratory environments, and experimental procedures by addressing the independence of data among pooled experiments. In our analysis, this was completed simply by including a discrete intercept covariate that indicated the source experiment in our models of the pooled data.

It is interesting that the time-to-event treatment effect *p*-values are smaller than corresponding problematic ANOVA estimates, but both are much smaller than logistic regression estimates. Although *p*-values are insufficient for formal model comparisons, it might be argued from an information theoretic perspective that because time-to-event models utilize information from both swim duration and incidence of finding the platform, such models are potentially more powerful than either ANOVA or logistic regression models, which use only a portion of that information. The theoretical basis for this claim relies on an assumption of consistency of these pieces of information in support of any explanatory hypothesis.

There is no question that the ANOVA type of model is fundamentally an incorrect model for MWM latency data due to potential bias in the presence of informative censoring. A logistic regression model for incidence of reaching the hidden platform is unaffected by censoring but has relatively low power due to the limited amount of information modeled. Time-to-event models would therefore appear to be the optimal means of modeling MWM latency, because they take advantage of information from both swim duration and incidence of reaching the platform and correctly handle any potential bias due to censoring.

The choice of which type of time-to-event model to use should be driven by the nature of the experiment.[29] Data from a single swim trial with two treatment groups might be adequately modeled by a Kaplan-Meier plot and log-rank test; however, a single swim trial is unlikely with MWM studies. Given multiple swims per animal, possibly over multiple days, or the need to adjust for prognostic covariates, a Cox proportional hazards model or an AFT model would be a better modeling choice, allowing for appropriate accommodation of repeated measures per animal among swims through clustering. Although Cox models are popular, the proportional hazards assumption required by those models may be violated in the context of MWM, as we found in this study. (Note that we had sufficient data to clearly detect this proportional hazards violation; studies with smaller sample size may not be powerful enough to detect such a violation.) Additionally, the interpretation of hazard ratios presents a challenge. AFT models have capabilities similar to those of Cox models in the MWM context, yet do not require proportional hazards, and have the added benefit of intuitive interpretation of estimates in terms of ratios between mean durations. These considerations lead us to suggest that AFT models should be the preferred analytical methodology for latency to platform outcomes associated with MWM studies.

Possible future directions to consider are the impact of using AFT models for MWM data obtained using different rat strains, females, and rats that are either juvenile or over 18 months of age. In a cortical contusion model of injury where MWM latency to platform was evaluated in adult, male, Sprague-Dawley rats and compared with age- and strain-matched female rats that were either proestrous or non-proestrous no sex-based differences were found.[30] We would not anticipate any differences in data analysis methodology for any of the other groups (age and strain); however, the logistics for the MWM testing itself may change with the physical capabilities of each group (a previous experience with aged male Sprague-Dawley rats required modifications to the platform for them to be able to climb onto it due to their increased weight with increased years of life).

The analyses reported here are not intended as templates for others to use blindly. Investigators should utilize models that are in accord with their study design and data. The applied analyses reported here have taken advantage of a combination of data from different experiments (of a specific and perhaps uncommon design), with a correspondingly large number of animals. This allowed the simultaneous modeling of all treatment days and swims, including a three-way interaction between day and swim and treatment, and facilitates reporting of treatment effects with an unusual level of precision. More typical studies would have much smaller numbers of animals, and would perhaps need to use less complex models, such as modeling the data from each day separately—which would result in the limitation of being unable to test for between-day differences.[29] Additionally, the statistical complexity associated with mixed-effect and time-to-event models may hopefully motivate collaboration with professional statisticians to take advantage of their specialized expertise when designing research studies.

Acknowledgments

We thank our research staff (Ian Bolding, Nicole Hausser, Kathia Johnson, and Margaret Parsley) who conducted the experiments used in this article and Andrew Hall for editorial assistance. We also acknowledge The Moody Project for Translational Traumatic Brain Injury Research team for thoughtful discussions and sharing data.

Funding Information

Authors were supported in part by the Darrell K. Royal Research Fund for Alzheimer's Disease (B.E.H.); Mission Connect, a Program of TIRR (B.E.H.); The Moody Project for Translational TBI Research (D.S.P.); and a National Institutes of Health BD2K TCC Data Science RoAD-Trip Fellowship (to B.E.H. from Parent Grant 5U24ES026465-02 to John Darrell VanHorn PI).

Author Disclosure Statement

No competing financial interests exist.

Supplementary Material

Supplementary Modeling S1
Supplementary Modeling S2
Supplementary Data Set S1
Supplementary Data Set S2
Supplementary Data Set S3

References

1. Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. J. Neurosci. Methods 11, 47–60.
2. Hamm, R.J., Dixon, C.E., Gbadebo, D.M., Singha, A.K., Jenkins, L.W., Lyeth, B.G., and Hayes, R.L. (1992). Cognitive deficits following traumatic brain injury produced by controlled cortical impact. J. Neurotrauma 9, 11–20.
3. Tucker, L.B., Velosky, A.G., and McCabe, J.T. (2018). Applications of the Morris water maze in translational traumatic brain injury research. Neurosci. Biobehav. Rev. 88, 187–200.
4. DeWitt, D.S., Hawkins, B.E., Dixon, C.E., Kochanek, P.M., Armstead, W., Bass, C.R., Bramlett, H.M., Buki, A., Dietrich, W.D., Ferguson, A.R., Hall, E.D., Hayes, R.L., Hinds, S.R., LaPlaca, M.C., Long, J.B., Meaney, D.F., Mondello, S., Noble-Haeusslein, L.J., Poloyac, S.M., Prough, D.S., Robertson, C.S., Saatman, K.E., Shultz, S.R., Shear, D.A., Smith, D.H., Valadka, A.B., VandeVord, P., and Zhang, L. (2018). Pre-clinical testing of therapies for traumatic brain injury. J. Neurotrauma 35, 2737–2754.
5. Hawkins, B.E., Huie, J.R., Almeida, C., Chen, J., and Ferguson, A.R. (2019). Data dissemination: shortening the long tail of traumatic brain injury dark data. J. Neurotrauma. DOI: 10.1089/neu.2018.6192 [Online ahead of print].
6. Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Stat. Assoc. 53, 457–481.
7. Cox, D.R. (1972). Regression models and life-tables. J. Royal Stat. Soc. 34, 187–220.
8. Jahn-Eimermacher, A., Lasarzik, I., and Raber, J. (2011). Statistical analysis of latency outcomes in behavioral experiments. Behav. Brain Res. 221, 271–275.
9. Tucker, L.B., Velosky, A.G., and McCabe, J.T. (2018). Applications of the Morris water maze in translational traumatic brain injury research. Neurosci. Biobehav. Rev. 88, 187–200.
10. Browne, K.D., Iwata, A., Putt, M.E., and Smith, D.H. (2006). Chronic ibuprofen administration worsens cognitive outcome following traumatic brain injury in rats. Exp. Neurol. 201, 301–307.
11. Jenks, K.R., Lucas, M.M., Duffy, B.A., Robbins, A.A., Gimi, B., Barry, J.M., and Scott, R.C. (2013). Enrichment and training improve cognition in rats with cortical malformations. PLoS One 8, e84492.
12. Boone, D.K., Weisz, H.A., Bi, M., Falduto, M.T., Torres, K.E.O., Willey, H.E., Volsko, C.M., Kumar, A.M., Micci, M.A., Dewitt, D.S., Prough, D.S., and Hellmich, H.L. (2017). Evidence linking microRNA suppression of essential prosurvival genes with hippocampal cell death after traumatic brain injury. Sci. Rep. 7, 6645.
13. Dewitt, D.S., Perez-Polo, R., Hulsebosch, C.E., Dash, P.K., and Robertson, C.S. (2013). Challenges in the development of rodent models of mild traumatic brain injury. J. Neurotrauma 30, 688–701.
14. Sell, S.L., Johnson, K., DeWitt, D.S., and Prough, D.S. (2017). Persistent behavioral deficits in rats after parasagittal fluid percussion injury. J. Neurotrauma 34, 1086–1096.
15. Hausser, N., Johnson, K., Parsley, M.A., Guptarak, J., Spratt, H., and Sell, S.L. (2018). Detecting behavioral deficits in rats after traumatic brain injury. J. Vis Exp. DOI: 10.3791/56044.
16. Nichols, J.E., Niles, J.A., DeWitt, D., Prough, D., Parsley, M., Vega, S., Cantu, A., Lee, E., and Cortiella, J. (2013). Neurogenic and neuro-protective potential of a novel subpopulation of peripheral blood-derived CD133+ ABCG2+CXCR4+ mesenchymal stem cells: development of autologous cell-based therapeutics for traumatic brain injury. Stem Cell Res. Ther. 4, 3.
17. Therneau, T.M., and Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag: New York.
18. Muller, H.-G., and Wang, J.L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. Biometrics 50, 61–76.
19. Grambsch, P.M., and Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81, 515–526.
20. Kalbfleisch, J.D., and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. John Wiley & Sons: Hoboken, NJ.
21. Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75, 383–386.
22. R Core Team. (2019). R: a language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org/ (Last accessed March 16, 2020).

23. Therneau, T.M., Lumley,T., Atkinson E., and Crowson C. (2020). Survival: survival analysis. Comprehensive R Archive Network. https://cran.r-project.org/package=survival (Last accessed March 16, 2020).
24. Hess, K. Gentleman, R., and Winsemius, D. (2019). muhaz: hazard function estimation in survival analysis. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/muhaz/ (Last accessed March 16, 2020).
25. Lenth, R., Singmann, H., Lobe, J., Buerkner, P., and Herve, M. (2020). emmeans: estimated marginal means, aka least-squares means. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/emmeans/index.html (Last accessed March 16, 2020).
26. Cumming, G. (2014). The new statistics: why and how. Psychol. Sci. 25, 7–29.
27. Andersen, C.R. (2019). catseyes: create catseye plots illustrating the normal distribution of the means. Comprehensive R Archive Network. https://rdrr.io/cran/catseyes/ (Last accessed March 16, 2020).
28. Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.
29. Andersen, C.R., Hawkins, B.E., and Prough, D.S. (2019). Finding the hidden (statistical) platform. Crit. Care Med. 47, 480–483.
30. Wagner, A.K., Willard, L.A., Kline, A.E., Wenger, M.K., Bolinger, B.D., Ren D., Zafonte, R.D., and Dixon, C.E. (2004). Evaluation of estrous cycle stage and gender on behavioral outcome after experimental traumatic brain injury. Brain Res. 998, 113–121.

Address correspondence to:
Bridget E. Hawkins, PhD, MBA
Moody Project for Translational Traumatic Brain
Injury Research
Department of Anesthesiology
University of Texas Medical Branch
Galveston, TX 77555
USA

E-mail: behawkin@utmb.edu