Research Article

# A Rational Approach of Early Humane Endpoint Determination in a Murine Model for Cholestasis

*Xianbin Zhang[1#], Simone Kumstel[1#], Guanglin Tang[1#], Steven R. Talbot[2], Nico Seume[1], Kerstin Abshagen[1], Brigitte Vollmar[1] and Dietmar Zechner[1]*

[1]Rudolf-Zenker-Institute of Experimental Surgery, University Medical Center Rostock, Rostock, Germany; [2]Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany

**Abstract**

Reduction of animal suffering during *in vivo* experiments is usually ensured by continuously monitoring the health status using a score sheet and by applying humane endpoints. However, most studies do not evaluate the plausibility of score sheets and do not attempt to reduce the suffering of animals by determining earlier and, therefore, more humane endpoints. The present study uses data from BALB/cANCrl mice after bile duct ligation to retrospectively analyze which score sheet criteria are informative to determine humane endpoints. The performance of each single as well as combinations of multiple animal welfare parameters was analyzed by a Cox proportional-hazards model followed by Harrell's concordance index. The addition of behavioral parameters, such as burrowing activity, helped to define a more humane early endpoint for euthanizing these animals. Using this approach, we determined that a body weight loss of 10-20% combined with a reduction of burrowing activity by more than 79.4% was able to predict that these animals would die within two days. Thus, this approach successfully determined an earlier humane endpoint and will reduce the suffering of animals in future experiments. Application of such an approach or similar methods can contribute to the refinement of various animal experiments.

## 1 Introduction

According to animal protection laws enacted by most nations (EU, 2010; Germany, 2013), high animal welfare standards are a prerequisite to obtain permission to perform animal-based research. Moreover, these standards also provide an important foundation for high-quality biomedical research (Bayne and Würbel, 2014; Carbone and Austin, 2016). Thus, it is in the interest of the public and of the scientific community to alleviate the suffering of animals used for scientific purposes.

One key aspect of reducing animal suffering is to determine humane endpoints for timely euthanasia. Accepted criteria for humane endpoints are, for example, 20% body weight loss (Morton, 2000) and hypothermia or lethargy (Acred et al., 1994). However, these symptoms often only occur under severe suffering just before death. Defining criteria that are able to predict death at an earlier time point could reduce the suffering of laboratory animals. A clinical score sheet or "welfare assessment protocol" was established by Morton and Griffiths (1985) as a tool to grade the suffering of animals and to determine humane endpoints. The score sheet should include reasonable criteria to recognize pain, suffering, or discomfort of the animals. Common criteria are body weight loss, appearance, spontaneous and flight behavior as well as intervention-specific clinical signs. According to these criteria, the distress of an animal can be classified as mild, moderate or severe (Morton and Griffiths, 1985; Hawkins et al., 2011; Smith et al., 2018).

Many different scoring systems have been established for different animal models (Paster et al., 2009; Kanzler et al., 2016), and animal welfare organizations have published many helpful protocols to improve the score sheets (Hawkins et al., 2011; Smith et al., 2018). These organizations also recommend the use of behavioral parameters to analyze the psychological state of the animals in addition to physical criteria such as body weight, posture, body tem-

[#]contributed equally

Received September 11, 2019; Accepted November 27, 2019;
Epub December 9, 2019; © The Authors, 2020.

**ALTEX 37(2), 197-207.** doi:10.14573/altex.1909111

Correspondence: PD Dietmar Zechner, PhD
Rudolf-Zenker-Institute for Experimental Surgery
University Medical Center Rostock
Schillingallee 69a, 18057 Rostock, Germany
(dietmar.zechner@uni-rostock.de)

perature, etc. (Hawkins et al., 2011). Thus, well-being in rodents is also often assessed by analyzing natural behaviors such as burrowing and nesting activity (Deacon, 2006a,b; Jirkof et al., 2013b; Jirkof, 2014). This has proven to be useful to detect neurological, abdominal, and post-surgical pain or stress in mice and rats (Deacon et al., 2005; Jirkof et al., 2013a,b; Jirkof, 2014; Pfeiffenberger et al., 2015; Sliepen et al., 2019; Jirkof et al., 2010).

Although many different scoring systems to determine animal suffering have been published to provide criteria for humane euthanasia of experimental animals, very few publications have attempted to define which read-out parameters are actually informative (Nemzek et al., 2004; Mai et al., 2018; Leung et al., 2019), and even fewer publications describe methods to optimize their scoring system in order to determine early humane endpoints (Nunamaker et al., 2013; Koch et al., 2016).

This study assesses multiple animal welfare parameters on mice who underwent bile duct ligation (BDL), a widely used procedure to study liver damage and fibrosis. We critically evaluate a scoring system in order to explore whether it can be optimized by measuring animal behavior and to determine more humane, earlier endpoint criteria.

## 2 Animals, material and methods

*Animals*

For this study we used a total of 55 male, 10/9.6-13.1-week-old (median/interquartile range) BALB/cANCrl mice with an average body weight of 25.6/23.9-27.4 g (median/interquartile range). Breeding pairs were originally purchased from Charles River and bred in the facility of the University Medical Center in Rostock under specific pathogen-free conditions. The mice had an acclimatization time of more than 2 days before the experiments started. During the experiment, the mice were kept in Eurostandard Type III plastic cages (Zoonlab GmbH, Castrop-Rauxel, Germany) with a light-dark cycle of 12 h/12 h at a temperature of 21 ±2°C (dawn: 6:30-7:00 a.m.) and a relative humidity of 60 ±20%. Food (pellets, 10 mm, ssniff-*Spezialdiäten* GmbH, Soest, Germany) and tap water were provided *ad libitum*. Enrichment was supplied by shredded tissue paper (PZN03058052, FSMED *Verbandmittel* GmbH, Frankenberg, Germany) as nesting material, one paper tunnel (75 × 38 mm, H 0528-151, ssniff), and a wooden enrichment tool (Espe size S, 40 × 16 × 10 mm, H0234.NSG, Abedd, Vienna, Austria). Due to low sociability and high aggression of male BALB/cANCrl mice (Brodkin, 2007; Jones and Brain, 1987), the animals were single-housed during the experiments. The animal experiment was approved by the local ethics committee and public authority (*Landesamt für Landwirtschaft, Lebensmittelsicherheit und Fischerei Mecklenburg-Vorpommern*, 7221.3-1-002/17), in accordance with European Directive 2010/63/EU (EU, 2010) as well as the national law of Germany, and is reported according to the ARRIVE Guidelines (Kilkenny et al., 2010).

*Induction of liver damage*

For the induction of cholestasis by BDL, mice were anesthetized by 1.2-2.5 vol. % isoflurane (CP-pharma, Burgdorf, Germany) and placed on a heating plate at 37°C in the laboratory. Isoflurane was chosen because it allows a fast recovery from anesthesia. 5 mg/kg carprofen (Rimadyl®, Pfizer GmbH, Berlin, Germany) was injected subcutaneously 5-15 min before the start of the surgical intervention for perioperative analgesia. The eyes were kept wet with eye ointment. A midline laparotomy was performed and the bile duct was ligated three times with 5-0 silk and transected between the two distal ligations (Abshagen et al., 2015). The peritoneum and the skin were closed separately with 6-0 and 4-0 prolene suture (Johnson & Johnson MEDICAL GmbH, New Brunswick), and the mice were placed in front of a heating lamp. The surgical procedure lasted 25-40 min. Wet pellets (10 mm, ssniff-*Spezialdiäten* GmbH) were provided as refinement during the first days of recovery. 1250 mg/L metamizol (Ratiopharm, Ulm,

**Tab. 1: Score-Sheet**
Kumstel et al., 2019

| Observations (variables) | Score |
|---|---|
| **I Body weight** | |
| I-a decreased > 10% (compared to initial weight) | 2 |
| I-b decreased > 20% (compared to initial weight) | 5 |
| **II General condition** | |
| II-a tooth displacement, too long teeth | 1 (A) |
| II-b fur dull, ruffled or untended | 2 |
| II-c eyes unclear or squinted | 2 |
| II-d untended orifices of the body | 3 |
| II-e abnormal posture | 3 |
| II-f dehydration | 3 |
| II-g short spasms or temporary paralysis symptoms | 3 |
| II-h persistent (>30') cramping or paralysis | 5 |
| II-i abnormal respiratory sounds or animal feels cold | 5 |
| **III Spontaneous behavior** | |
| III-a the animal is passive or overactive | 2 |
| III-b pronounced apathy, hyperkinetic, or isolation | 4 |
| III-c squeaking due to pain | 5 |
| III-d self-mutilation | 5 |
| **IV Flight behavior after contact** | |
| IV-a animal is passive or overactive | 2 |
| IV-b distinct apathy or hyperkinetic | 5 |
| **V Process-specific criteria** | |
| V-a wound healing disorder | 2 |
| V-b opening of the sutures by biting | 1 (B) |
| V-c local inflammation | 2 |
| V-d ascites | 4 |
| Total score | 0-66 |

Germany) was administered via the drinking water through the whole experiment for pain relief/management. In order to evaluate a possible therapeutic efficacy of the NLRP3 inflammasome inhibitor MCC950 (Sigma Aldrich, St. Louise, USA), 20 mg/kg MCC950 or aqua (control) was injected daily between 8:00-10:00 a.m. intraperitoneally from day -1 before BDL to day 13 after BDL. Animals were allocated in a non-random manner, matching the age of both treatment groups, and the researchers were not blinded when injecting drugs. The mice were euthanized by cervical dislocation after a short anesthesia by 5 vol. % isoflurane on day 14 after BDL or when one of the humane endpoint criteria was met according to the score sheet (Tab. 1).

*Cohorts of mice and assessment of distress*

Of a total of 55 mice, 20 were euthanized or did not survive until day 14 after BDL. These mice were defined as non-survivors and were 10.0/8.9-12.1 weeks old (median/interquartile range) at the beginning of the experiment and had a body weight of 25.0/23.8-26.6 g (median/interquartile range). The other 35 mice survived until day 14 after BDL and were therefore defined as survivors. The survivors had an age of 11.9/10.1-13.7 (median/interquartile range) weeks and a weight of 26.7/24.3-27.8 g (median/interquartile range) at the start of the experiment. 75% (15/20) of non-survivors were euthanized or died within 4 days after BDL. These mice died within 3.0/2.0-4.75 (median/interquartile range) days.

We assessed distress on day 1 and day 4 after BDL in the survivor and non-survivor cohort. For those non-survivors from which we could not obtain the data on day 4, we adopted the data points measured 0-2 days before death or euthanasia. The distress of all 55 mice (35 survivors = 70 data points; 20 non-survivors = 33 data points since 7 mice had to be euthanized before the second score could be assessed; 103 data points in total) was evaluated using a score sheet (Tab. 1 and Tab. S1[1]). The score-sheet was based on previously published score sheets (Morton and Griffiths, 1985; Paster et al., 2009) and had already been used in our group to evaluate murine animal models for gastrointestinal diseases (Kumstel et al., 2019). The distress score was assessed between 8:30-10:30 a.m. in the home cage. According to the defined score sheet, body weight, appearance, spontaneous and flight behavior as well as intervention-specific clinical signs were assessed in a non-blinded fashion by two observers (GT, NS), and in case of discrepancies by a third observer (DZ).

Burrowing behavior (variable VI) of 24 mice (16 survivors = 32 data points; 8 non-survivors = 13 data points, 3 mice had to be euthanized before the second data point was assessed; 45 data points in total) was analyzed according to Deacon (2006b). A tube was filled with 200 g pellets (10 mm, ssniff-*Spezialdiäten* GmbH) and placed into the home cage 2.5-3 h before the dark phase (at 16:00-16:30); the burrowed amount of pellets was calculated 17 h later.

Nesting activity (variable VII) was assessed on a different cohort of animals than burrowing activity, as the two assessments might influence each other. Nesting activity was evaluated on 22 mice (15 survivors = 30 data points; 7 non-survivors = 10 data points since 4 mice did not survive long enough; 40 data points in total). Nesting activity was assessed by providing a nestlet in the home cage (5 cm square of pressed cotton batting, Zoonlab GmbH, Castrop-Rauxel, Germany) 0.5-1 h before the dark phase (at 18:00-18:30). The nest was scored the next morning (at 9:00-11:00 a.m.) according to the 1-5 point scale of Deacon (2006a). We additionally scored 6 points for a perfect nest when more than 90% of the circumference of the walls was higher than the mouse. To enable individual learning, both behavior tests were performed three times in group housing before the mice were housed separately.

Body weight reduction was assessed in 48 mice (35 survivors = 70 data points; 13 non-survivors = 21 data points since 5 mice died before the second data point for bodyweight reduction could be assessed; 91 data points in total).

*Development of an optimal prognosis model*

Continuous variables, such as reduction of burrowing and nesting activity, were converted into dichotomous variables according to Youden's index (Ruopp et al., 2008). Based on this approach, the best cut-off value to distinguish between survivors and non-survivors was at 79.4% reduction of burrowing behavior or a nesting score of less than 2.5. The Kaplan-Meier estimator followed by log-rank test was performed by SigmaPlot 12.0 (SYSTAT Software Inc., San Jose, USA), and all variables that could significantly discriminate between the survival time of non-survivors and survivors were used to develop the prognosis model by univariate and multivariate Cox proportional-hazards model (SigmaPlot 12.0, SYSTAT Software Inc., San Jose, USA). To evaluate whether the proportional hazards assumption is satisfied, we performed log-minus-log plots, a frequently used method for the validation of a proportional hazard assumption (In and Lee, 2018), for all variables used for strategies 1, 2 and 3. To determine the optimal prognosis model, Harrell's concordance indices (C-indices) were investigated using the Hmisc (Harrell, 2019) and survival packages (Liu, P. et al., 2017; Therneau, 2019) of R software (Foundation for Statistical Computing, Vienna, Austria). C-indices were internally validated by 10'000-fold bootstrapping of the parametrized survival models using the boot package (Canty and Ripley, 2019; Davison and Hinkley, 1997). Resulting means and the bias-corrected and accelerated (BCa) 95% confidence intervals (CI) are reported.

## 3 Results

### 3.1 Retrospective analysis of distress parameters

After removing distress parameters that were never observed during the experiment (II-a, II-d, II-g, II-h, III-c, III-d, V-a, V-b, V-c) and parameters that demanded immediate euthanasia (I-b, II-i, IV-b), 11 variables were evaluated by Kaplan-Meier estimator (Fig. 1). There was no significant difference regarding the survival time when considering some single variables, such as criteria II-b (fur dull, ruffled or untended), II-e (abnormal pos-

[1] doi:10.14573/altex.1909111s

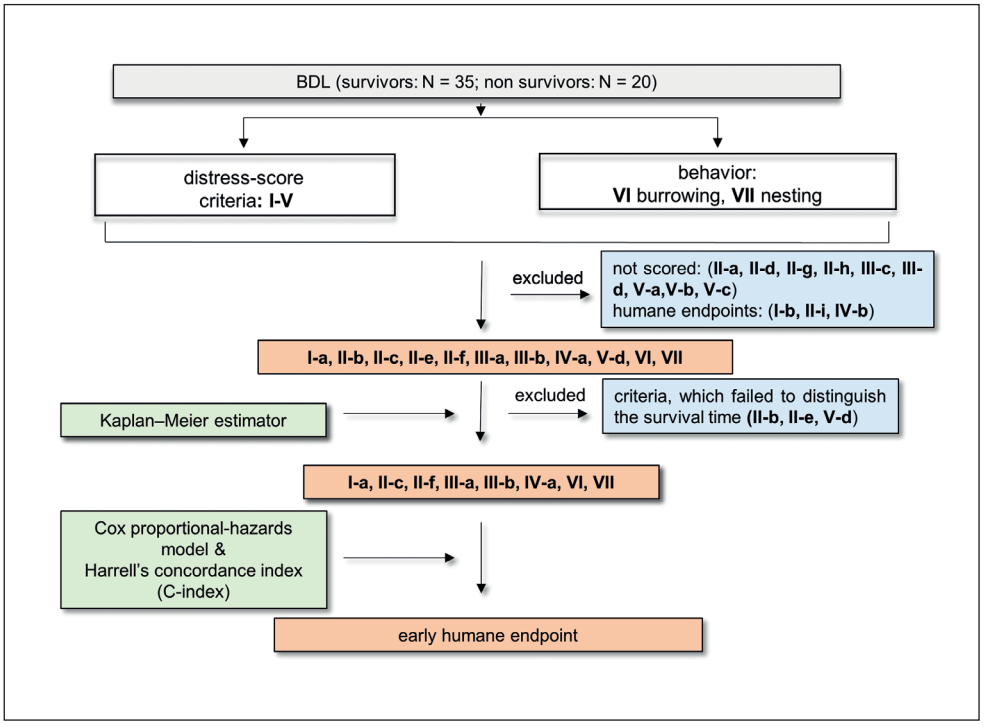



**Fig. 1: Flowchart to retrospectively analyze distress parameters in order to determine an early humane endpoint**
First, score sheet criteria (for details see Tab. 1) that were not observed during the experiment as well as humane endpoint criteria were excluded. Second, Kaplan-Meier estimator curves were used to exclude criteria that did not predict survival time. Third, the performance of each single as well as combinations of multiple parameters were analyzed by Cox proportional-hazards model followed by Harrell's concordance index to determine which criteria combination might be used as an efficient early humane endpoint.

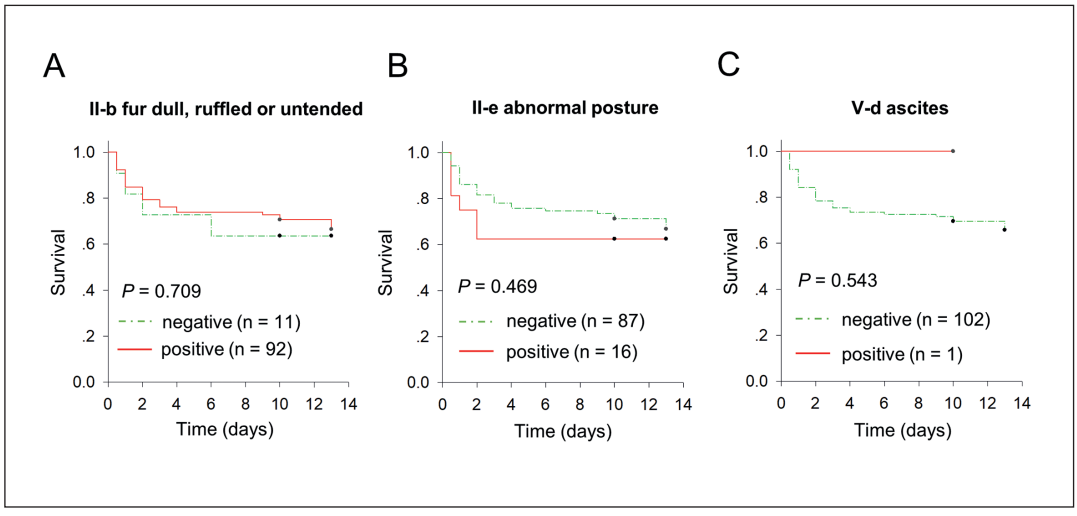



**Fig. 2: Kaplan-Meier curves using distinct score sheet criteria**
The positive status of II-b (fur dull, ruffled or untended) (A), II-e (abnormal posture) (B) or V-d (ascites) (C) failed to significantly indicate reduced survival time of mice. The *P*-value was determined by log-rank test.

ture) and V-d (ascites) (Fig. 2). However, mice with a positive status of I-a (body weight loss of 10% to 20%), II-c (eyes unclear or squinted), II-f (dehydration), III-a (spontaneous behavior: passive or overactive), III-b (pronounced apathy, hyperkinetic, or isolation), IV-a (passive or overactive after being touched by the observer), VI (reduction of burrowing activity more than 79.4%), or VII (decrease of nesting score more than 2.5) had a significantly ($P \leq 0.005$) shorter survival time than mice with negative status of these criteria (Fig. 3). All mice with a positive status of I-a had to be euthanized within 14 days, and mice with pronounced apathy (III-b) or who were passive or overactive after being touched by the researcher (IV-a) had to be euthanized within 1-2 days. This suggests that I-a, III-b and IV-a might be powerful criteria to define early humane endpoints for the BDL model.

### 3.2 Predictive models for defining survival time

We first used a univariate Cox proportional-hazards model in order to evaluate the hazard ratio (HR) of each variable. A positive status of I-a, II-c, II-f, III-a, III-b, IV-a, VI or VII significantly increased the risk of death when compared to the negative status of these variables (Tab. 2). We then performed a multivariate analysis of these distress score criteria (strategy 1). Two vari-

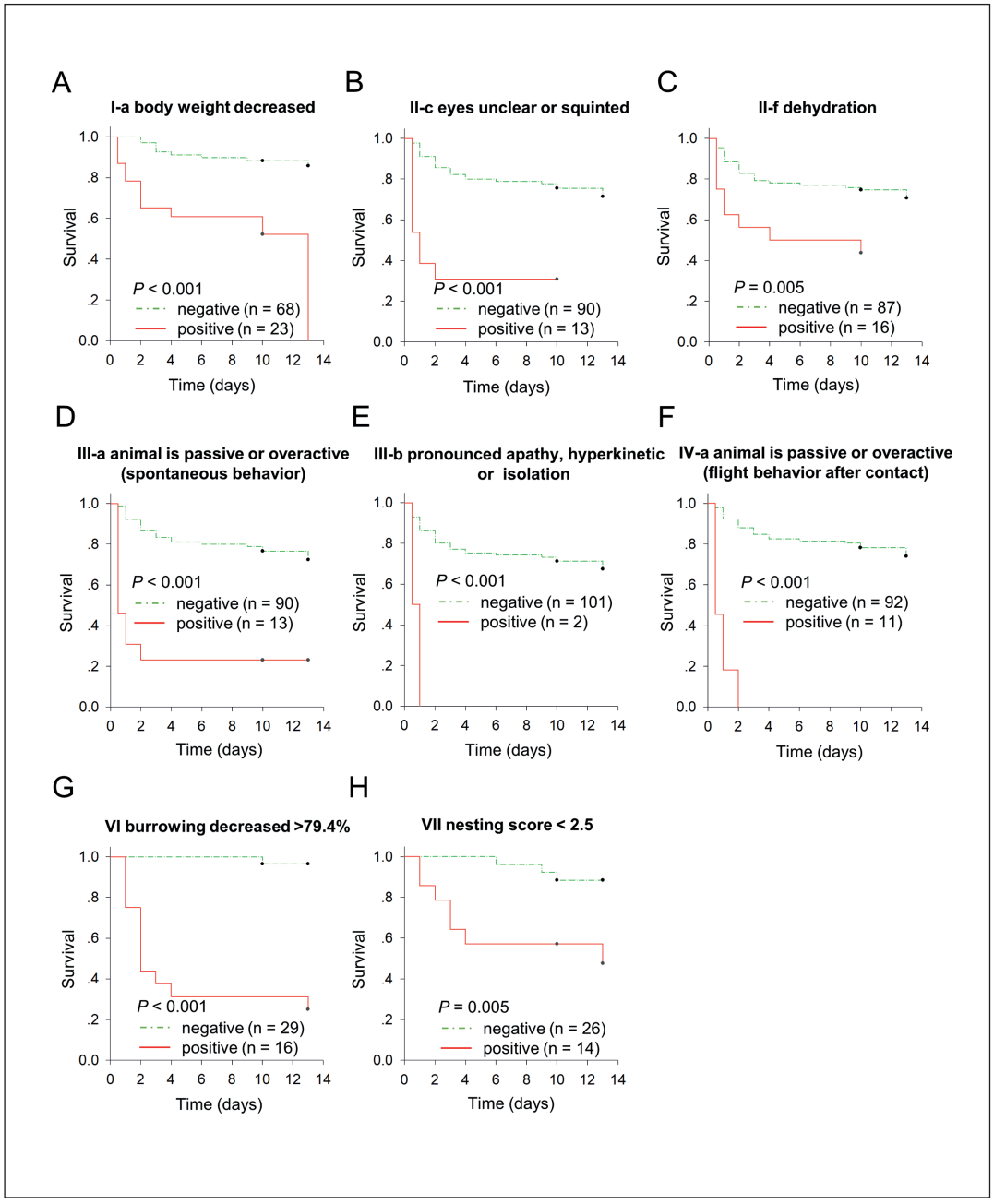



**Fig. 3: Kaplan-Meier curves using distinct read-out parameters for distress**

For the positive status of I-a (body weight decreased 10 to 20%) (A), II-c (eyes unclear and squinted) (B), II-f (dehydration) (C), III-a (spontaneous behavior: animal is passive and over active) (D), III-b (pronounced apathy, hyperkinetic or isolation) (E), IV-a (flight behavior after contact: animal is passive or overactive;) (F), decreased burrowing activity by > 79.4% (G) and nesting score of less than 2.5 (H), a significantly shorter survival time was calculated by log rank test ($P \leq 0.005$).

ables, I-a and IV-a, were found to significantly affect the hazard rate. Thus, a positive status of these two variables significantly increased the risk of death (strategy 1 in Tab. 2).

We then included two behavioral parameters in our analysis, i.e., burrowing and nesting activity. Since these two variables were assessed in two distinct cohorts, we developed two additional multivariate Cox proportional-hazards models (strategy 2 and strategy 3). First, we added burrowing activity to all distress score variables (strategy 2). We observed that the positive status of variable I-a (body weight) and VI (burrowing) significantly increased the risk of death (strategy 2 in Tab. 2).

We also added nesting activity to all distress score variables (strategy 3). We observed that a multivariate Cox proportional-hazards model only recognized that nesting activity significantly increased the risk of death (strategy 3 in Tab. 2).

The following observations support the concept that the conditions for applying the Cox proportional-hazard model are met: First, we do not observe that curves in the Kaplan-Meier curves cross. Second, we made log-minus-log plots for all variables used for strategy 1 (Fig. S1[1]), 2 (Fig. S2[1]) and 3 (Fig. S3[1], for strategies see Tab. 2) and observed that these curves are parallel.

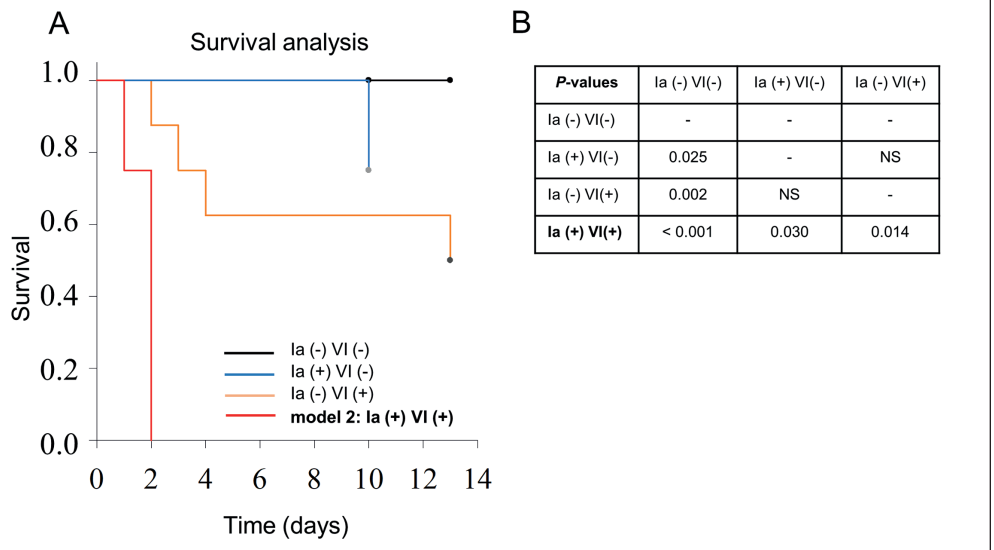
**Tab. 2: HR (hazard ratio) and *P*-value (P) for the distinct variables were determined by Cox proportional-hazards model**

| Variables | | Univariate | | Multivariate | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | HR | *P* | Strategy 1 | | Strategy 2 | | Strategy 3 | |
| | | | | HR | *P* | HR | *P* | HR | *P* |
| I-a | negative | reference | | reference | | reference | | NS | |
| | positive | 6.169 | < 0.001 | 3.968 | 0.017 | 24.096 | 0.007 | | |
| II-c | negative | reference | | NS | | NS | | - | |
| | positive | 5.033 | < 0.001 | | | | | | |
| II-f | negative | reference | | NS | | NS | | NS | |
| | positive | 2.825 | 0.009 | | | | | | |
| III-a | negative | reference | | NS | | NS | | NS | |
| | positive | 6.584 | < 0.001 | | | | | | |
| III-b | negative | reference | | NS | | - | | - | |
| | positive | 9.495 | 0.003 | | | | | | |
| IV-a | negative | reference | | reference | | NS | | - | |
| | positive | 17.130 | < 0.001 | 8.621 | 0.021 | | | | |
| VI | negative | reference | | - | | reference | | - | |
| | positive | 33.218 | < 0.001 | | | 54.348 | 0.003 | | |
| VII | negative | reference | | - | | - | | reference | |
| | positive | 5.639 | 0.013 | | | | | 5.051 | 0.026 |

NS indicates no significant difference (P > 0.05).

**Tab. 3: Bootstrapped C-indices (Harrell's concordance index) and the corresponding 95% confidence intervals for each distinct model**

| | Variables | C-index | C-index bootstrapped | 95% CI |
|---|---|---|---|---|
| Single variables | I-a | 0.691 | 0.692 | 0.572-0.797 |
| | II-c | 0.633 | 0.632 | 0.553-0.721 |
| | II-f | 0.590 | 0.590 | 0.517-0.680 |
| | III-a | 0.657 | 0.656 | 0.574-0.746 |
| | III-b | 0.534 | 0.534 | 0.500-0.614 |
| | IV-a | 0.684 | 0.683 | 0.603-0.770 |
| | VI | 0.865 | 0.866 | 0.768-0.923 |
| | VII | 0.725 | 0.724 | 0.534-0.848 |
| Model-1 | Ia, IVa | 0.719 | 0.720 | 0.590-0.838 |
| **Model-2** | **Ia, VI** | **0.943** | **0.947** | **0.832-0.978** |
| Model-3 | VII | 0.725 | 0.724 | 0.532-0.849 |
| Model-4 | I-a, II-c, II-f, III-a, III-b, IV-a | 0.696 | 0.726 | 0.553-0.782 |




B

| *P*-values | Ia (-) VI(-) | Ia (+) VI(-) | Ia (-) VI(+) |
|---|---|---|---|
| Ia (-) VI(-) | - | - | - |
| Ia (+) VI(-) | 0.025 | - | NS |
| Ia (-) VI(+) | 0.002 | NS | - |
| **Ia (+) VI(+)** | < 0.001 | 0.030 | 0.014 |

**Fig. 4: Kaplan-Meier curve of the multivariate model 2**
Mice with a body weight loss of 10 to 20% (variable I-a) and decreased burrowing activity of more than 79.4% (variable VI) died within 2 days (A). Mice positive for variable Ia and VI had a significantly decreased survival time when compared to mice positive for just one variable or negative for both variables (B). The *P*-values were determined by log-rank test ($P < 0.05$).

### 3.3 Evaluation of models for defining survival time

Models that use single variables or combinations of variables were compared to each other using Harrell's concordance index. We bootstrapped the indices of each model to obtain more robust estimates, thereby internally validating the goodness-of-fit (Tab. 3) of each model. Model 2 (C-index: 0.947, 95% CI: 0.832-0.978) has a higher C-index than model 1 (C-index: 0.720, 95% CI: 0.590-0.838) or model 3 (C-index: 0.724, 95% CI: 0.532-0.849).

In addition, we evaluated model 4 (C-index: 0.726, 95% CI: 0.553-0.782), in which we used a combination of all distress score parameters that significantly increased the risk of death based on the univariate Cox proportional-hazards model. Model 2 also had a higher C-index than model 4 (Tab. 3).

The C-index of model 2, body weight plus burrowing activity (C-index: 0.947, 95% CI: 0.832-0.978), was also higher than that of each single variable (body weight C-index: 0.692, 95% CI: 0.572-0.797; burrowing activity weight C-index: 0.866, 95% CI: 0.768-0.923) as well as the C-index of all other single variables (Tab. 2). Note that a C-index of 0.5 corresponds to a non-informative prediction rule, whereas a C-index of 1 corresponds to a perfect prediction rule (Schmid et al., 2016). Since the C-index as well as the bootstrapped C-index of our model 2 is > 0.9, this suggests that model 2 can predict the survival times of these animals very well.

In order to evaluate the practicability of using this combination, we evaluated it by Kaplan-Meier estimator. We observed that mice who lost body weight in a range of 10% to 20% and also showed decreased burrowing activity by more than 79.4% had a significantly shorter survival time when compared to mice who were positive for only one or neither of these two variables (Fig. 4). All mice with such a reduction in body weight and burrowing activity died within 2 days (Fig. 4).

Lastly, we analyzed whether age influences the humane endpoint. We determined the optimal cut-off of age (10.21 weeks) by Youden's index and observed that young mice (age < 10.21 weeks) had a significant reduction of survival time when compared to older mice (Fig. S4A[1]). We compared how our best model (using variable I-a plus variable VI) predicts death in young or older mice and observed that this model could predict death of older mice within one day (Fig. S4B[1]) and death of young mice within two days (Fig. S4C[1]).

## 4 Discussion

In the current study, we present an approach of how to retrospectively analyze a score sheet and how to include additional distress parameters in order to define an early humane endpoint. Mice who lost 10-20% body weight and reduced their burrowing activity by more than 79.4% died within 2 days. Thus, applying these cut-off criteria for an earlier euthanasia could reduce the suffering of animals in subsequent experiments. Application of such an approach could contribute to the refinement of most animal experiments.

The approach was based on three steps (Fig. 1): First, score sheet criteria that were not observed at all during the experiment were excluded as well as criteria that called for immediate euthanasia. Second, Kaplan-Meier estimator curves helped to exclude criteria that did not contribute to predicting survival times. Third, the performance of each single as well as combinations of multiple parameters were analyzed by a Cox proportional-hazards model followed by Harrell's concordance index.

Interestingly, survival analysis using the Kaplan-Meier estimator suggested that the single variables III-b (pronounced ap-

athy, hyperkinetic or isolation) and IV-a (animal is passive or overactive- flight behavior after contact) are able to predict death within 1-2 days (Fig. 3E,F). Therefore, one could consider these two criteria to determine an early humane endpoint. However, the incidence of variable III-b (n = 1) was low and the accuracy according to the C-indices (III-b: 0.534; IV-a: 0.684) of these two criteria was also low compared to that of multivariate model 2 (C-index: 0.943). This indicates that a multivariate distress analysis can be superior to a univariate analysis. This conclusion is consistent with other studies that also suggest that using more than one parameter for death prediction is beneficial (Hankenson et al., 2013; Trammell and Toth, 2011; Ray et al., 2010).

Our multivariate approach was implemented by the Cox proportional-hazard model and Harrell's concordance index of models. These methods are commonly applied in statistical medical research to investigate the survival time of patients reliant on one or more independent variables. Using this approach, we were able to define distinct univariate and multivariate models. A comparison of the C-indices revealed that the combination of body weight loss and reduction of burrowing behavior is the best model for predicting survival times (Tab. 3). An internal validation of the goodness-of-fit via bootstrapping the C-indices of each model also demonstrated the robustness of model 2. Model 2 has a C-index well above 0.9. A value of C = 0.5 corresponds to a non-informative prediction rule, whereas C = 1 corresponds to a perfect prediction rule (Schmid et al., 2016). In many meaningful biomedical applications, Harrel's concordance index typically ranges between the values 0.6 and 0.75. This was reported, for example, by Van Belle et al. (2011), Schröder et al. (2011) and Zhang et al. (2013). Since both the C-index and the bootstrapped C-index of our model 2 is above 0.9, we conclude that model 2 can predict the survival times of these animals very well. The two variables body weight and burrowing behavior have the additional advantage that they can be objectively measured, which minimizes potential selection bias. It is also well-accepted that both criteria reliably measure suffering of mice (Jirkof et al., 2013b; Deacon et al., 2005; Pfeiffenberger et al., 2015; Hohlbaum et al., 2017; Häger et al., 2018). However, it is possible that for each animal model a different combination of parameters might be best to predict survival times and to determine early humane endpoints.

This study relied on the evaluation of physical parameters, i.e., score sheet criteria and body weight, or behavioral parameters, such as burrowing as well as nesting activity, for determining a humane endpoint. Body weight has been demonstrated to be very useful for determining humane endpoints in many other studies (Takayama-Ito et al., 2017; Trammell and Toth, 2011; Hankenson et al., 2013). The advantage of body weight is that it is applicable as a humane endpoint in many species, and it is easy to assess. However, for cancer studies this parameter proved to be less useful as an exclusive indicator for euthanasia (Paster et al., 2009). Since the adaption of body weight after an intervention or during a disease takes about 24 h, this endpoint criterion is also not practical for acute and severe diseases.

In contrast to body weight loss, only one unsuccessful attempt was published so far where burrowing activity was used to determine a humane endpoint in a chemotherapy-induced mucositis model in rats (Whittaker et al., 2015). However, burrowing behavior is known to be a sensitive indicator for distress after surgical interventions (Jirkof et al., 2010, 2013a) or during chronic diseases (Jirkof, 2013b; Abdelrahman et al., 2019). Even the suffering caused by neurological disorders (Deacon et al., 2005; Felton et al., 2005) or very mild stress inductions such as isoflurane aesthesia (Hohlbaum et al., 2017) can be quantified by burrowing behavior. A disadvantage of burrowing activity is that it is only applicable for rodents. Subjective score sheet criteria, i.e., cramping and paralysis, abnormal respiratory sounds, squeaking due to pain, self-mutilation or apathy and hyperkinetic behavior, also have been introduced in other studies as humane endpoints for mice and rats (Kanzler et al., 2016; Brabb et al., 2014; Herrmann and Flecknell, 2018).

Concerning bile duct ligation in mice, the following criteria were defined so far as humane endpoints: distension of abdomen, ascites, debilitating diarrhea, bleeding from the orifices, peritonitis, internal bleeding or sepsis (Tag et al., 2015a,b). These humane endpoints mostly describe pathologies instead of symptoms. The presented score sheet criteria focus on symptoms that might be caused by these pathologies.

Some studies used body temperature as an important parameter for determining humane endpoints (Mai et al., 2018; Mei et al., 2018; Warn et al., 2003; Drechsler et al., 2015; Kort et al., 1998; Cates et al., 2014). Body temperature is considered to be the most accurate humane endpoint criterion, especially for acute and severe infections (Nemzek et al., 2004; Adamson et al., 2013). Thus, the lack of an accurate measurement of body temperature might be a limitation of this study. A major drop in body temperature is observed in sepsis models and acute infections a few hours before death (Napier et al., 2016; Mai et al., 2018; Kort et al., 1998). Since it was our goal to develop early humane endpoints, we abstained from measuring the body temperature.

Another limitation of this study might be that we determined an earlier humane endpoint by using only data from day 1 and day 4 after BDL. To address this point, we added additional data taken at day 2 (from 8 non-survivors and 35 survivors) and analyzed survival by Kaplan Meyer curves. This resulted in very similar survival curves as shown in Figure 3, suggesting that additional data have little influence on our results (data not shown). Since we could demonstrate that older mice (> 10.21 weeks) have a better survival than young mice, using older mice will reduce the number of animals needed and might help to reduce the distress caused by this animal model.

Some studies have also used multivariate analysis for early humane endpoint determination (Trammell and Toth, 2011; Nunamaker et al., 2013). However, these studies did not compare the accuracy of single criteria or multiple combinations for predicting death. By using the combination of up to three parameters, i.e., body weight, temperature and a neuroscore, an elaborate machine learning approach for early humane endpoint determination in mouse models for stroke and sepsis has been established (Mei et al., 2019). The accuracy of this model (stroke: 0.93; sepsis: 0.96) proved to be quite high. However, the machine learning tool needs to be trained with physical data from the specific animal model and a large sample size is necessary to build a generalized

model. Hankenson et al. established an early humane endpoint model using the combination of body temperature (< 34.5°C) and weight loss (more than 0.05 g daily) by linear regression (Hankenson et al., 2013). Linear regression estimates diagnostic outcomes at the moment of prediction, while Cox regression determines prognostic outcomes within a distinct period of time (Moons et al., 2015). The inclusion of time as a factor is an advantage of our approach, because it also provides information on how fast an animal dies. Thus, compared to the already published methods for early humane endpoint determination, the advantage of our approach is the inclusion of the time variable by Cox regression and that a justifiable low number of animals is required for analysis.

## References


Abdelrahman, A., Kumstel, S., Zhang, X. et al. (2019). A novel multi-parametric analysis of non-invasive methods to assess animal distress during chronic pancreatitis. *Sci Rep 9*, 14084. doi:10.1038/s41598-019-50682-3

Abshagen, K., König, M., Hoppe, A. et al. (2015). Pathobiochemical signatures of cholestatic liver disease in bile duct ligated mice. *BMC Syst Biol 9*, 83. doi:10.1186/s12918-015-0229-0

Acred, P., Hennessey, T. D., MacArthur-Clark, J. A. et al. (1994). Guidelines for the welfare of animals in rodent protection tests. A report from the rodent protection test working party. *Lab Anim 28*, 13-18. doi:10.1258/002367794781065870

Adamson, T. W., Diaz-Arevalo, D., Gonzalez, T. M. et al. (2013). Hypothermic endpoint for an intranasal invasive pulmonary aspergillosis mouse model. *Comp Med 63*, 477-481.

Bayne, K. and Würbel, H. (2014). The impact of environmental enrichment on the outcome variability and scientific validity of laboratory animal studies. *Rev Sci Tech 33*, 273-280. doi:10.20506/rst.33.1.2282

Brabb, T., Carbone, L., Snyder, J. et al. (2014). Institutional animal care and use committee considerations for animal models of peripheral neuropathy. *ILAR J 54*, 329-337. doi:10.1093/ilar/ilt045

Brodkin, E. S. (2007). BALB/c mice: Low sociability and other phenotypes that may be relevant to autism. *Behav Brain Res 176*, 53-65. doi:10.1016/j.bbr.2006.06.025

Canty, A. and Ripley, B. (2019). boot: Bootstrap R (S-Plus) Funtions. R package version. Version 1.3-22.

Carbone, L. and Austin, J. (2016). Pain and laboratory animals: Publication practices for better data reproducibility and better animal welfare. *PLoS One 11*, e0155001. doi:10.1371/journal.pone.0155001

Cates, C. C., McCabe, J. G., Lawson, G. W. et al. (2014). Core body temperature as adjunct to endpoint determination in murine median lethal dose testing of rattlesnake venom. *Comp Med 64*, 440-447.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge, UK: Cambridge University Press (Cambridge series on statistical and probabilistic mathematics, 1). doi:10.1017/CBO9780511802843

Deacon, R. M. J., Reisel, D., Perry, V. H. et al. (2005). Hippocampal scrapie infection impairs operant DRL performance in mice. *Behav Brain Res 157*, 99-105. doi:10.1016/j.bbr.2004.06.013

Deacon, R. M. J. (2006a). Assessing nest building in mice. *Nat Protoc 1*, 1117-1119. doi:10.1038/nprot.2006.170

Deacon, R. M. J. (2006b). Burrowing in rodents: A sensitive method for detecting behavioral dysfunction. *Nat Protoc 1*, 118-121. doi:10.1038/nprot.2006.19

Drechsler, S., Weixelbaumer, K. M., Weidinger, A. et al. (2015). Why do they die? Comparison of selected aspects of organ injury and dysfunction in mice surviving and dying in acute abdominal sepsis. *ICMx 3*, 1247. doi:10.1186/s40635-015-0048-z

EU – The European Parliament and the Council of the European Union (2010). Directive 2010/63/EU of the European Parliament and of the Council of 22 of September 2010 on the protection of animals used for scientific purposes. 2010/63/EU. *Off J Eur Union L 276*, 33-79. http://data.europa.eu/eli/dir/2010/63/oj

Felton, L. M., Cunningham, C., Rankine, E. L. et al. (2005). MCP-1 and murine prion disease: Separation of early behavioural dysfunction from overt clinical disease. *Neurobiol Dis 20*, 283-295. doi:10.1016/j.nbd.2005.03.008

Germany (2013). Tierschutzgesetz, TierSchG, vom 04.07.2013 (BGBl.S1950). In *BGBl*. http://www.gesetze-im-internet.de/tierschg/BJNR012770972.html (accessed November 2019)

Häger, C., Keubler, L. M., Talbot, S. R. et al. (2018). Running in the wheel: Defining individual severity levels in mice. *PLoS Biol 16*, e2006159. doi:10.1371/journal.pbio.2006159

Hankenson, F. C., Ruskoski, N., van Saun, M. et al. (2013). Weight loss and reduced body temperature determine humane endpoints in a mouse model of ocular herpesvirus infection. *J Am Assoc Lab Anim Sci 52*, 277-285.

Harrell, F. (2019). Package 'Hmisc'. https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf (accessed August 2019)

Hawkins, P., Morton, D. B., Burman, O. et al. (2011). A guide to defining and implementing protocols for the welfare assessment of laboratory animals: Eleventh report of the BVAAWF/FRAME/RSPCA/UFAW Joint Working Group on Refinement. *Lab Anim 45*, 1-13. doi:10.1258/la.2010.010031

Herrmann, K. and Flecknell, P. (2018). Severity classification of surgical procedures and application of health monitoring strategies in animal research proposals: A retrospective review. *Altern Lab Anim 46*, 273-289. doi:10.1177/026119291804600606

Hohlbaum, K., Bert, B., Dietze, S. et al. (2017). Severity classification of repeated isoflurane anesthesia in C57BL/6JRj mice-assessing the degree of distress. *PLoS One 12*, e0179588. doi:10.1371/journal.pone.0179588

In, J. and Lee, D. K. (2018). Survival analysis: Part I – Analysis of time-to-event. *Korean J Anesthesiol 71*, 182-191. doi:10.4097/kja.d.18.00067

Jirkof, P., Cesarovic, N., Rettich, A. et al. (2010). Burrowing behavior as an indicator of post-laparotomy pain in mice. *Front Behav Neurosci 4*, 165. doi:10.3389/fnbeh.2010.00165

Jirkof, P., Fleischmann, T., Cesarovic, N. et al. (2013a). Assessment of postsurgical distress and pain in laboratory mice by nest complexity scoring. *Lab Anim 47*, 153-161. doi:10.1177/0023677213475603

Jirkof, P., Leucht, K., Cesarovic, N. et al. (2013b). Burrowing is a sensitive behavioural assay for monitoring general wellbeing during dextran sulfate sodium colitis in laboratory mice. *Lab Anim 47*, 274-283. doi:10.1177/0023677213493409

Jirkof, P. (2014). Burrowing and nest building behavior as indicators of well-being in mice. *J Neurosci Methods 234*, 139-146. doi:10.1016/j.jneumeth.2014.02.001

Jones, S. E. and Brain, P. F. (1987). Performances of inbred and outbred laboratory mice in putative tests of aggression. *Behav Genet 17*, 87-96. doi:10.1007/BF01066013

Kanzler, S., Rix, A., Czigany, Z. et al. (2016). Recommendation for severity assessment following liver resection and liver transplantation in rats: Part I. *Lab Anim 50*, 459-467. doi:10.1177/0023677216678018

Kilkenny, C., Browne, W. J., Cuthill, I. C. et al. (2010). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol 8*, e1000412. doi:10.1371/journal.pbio.1000412

Koch, A., Gulani, J., King, G. et al. (2016). Establishment of early endpoints in mouse total-body irradiation model. *PLoS One 11*, e0161079. doi:10.1371/journal.pone.0161079

Kort, W. J., Hekking-Weijma, J. M., TenKate, M. T. et al. (1998). A microchip implant system as a method to determine body temperature of terminally ill rats and mice. *Lab Anim 32*, 260-269. doi:10.1258/002367798780559329

Kumstel, S., Tang, G., Zhang, X. et al. (2019). Grading distress of different animal models for gastrointestinal diseases based on plasma corticosterone kinetics. *Animals 9*, 145. doi:10.3390/ani9040145

Leung, V. S. Y., Benoit-Biancamano, M.-O. and Pang, D. S. J. (2019). Performance of behavioral assays: The rat grimace scale, burrowing activity and a composite behavior score to identify visceral pain in an acute and chronic colitis model. *Pain Rep 4*, e718. doi:10.1097/PR9.0000000000000712

Liu, P., Zhang, X., Shang, Y. et al. (2017). Lymph node ratio, but not the total number of examined lymph nodes or lymph node metastasis, is a predictor of overall survival for pancreatic neuroendocrine neoplasms after surgical resection. *Oncotarget 8*, 89245-89255. doi:10.18632/oncotarget.19184

Mai, S. H. C., Sharma, N., Kwong, A. C. et al. (2018). Body temperature and mouse scoring systems as surrogate markers of death in cecal ligation and puncture sepsis. *Intensive Care Med Exp 6*, 20. doi:10.1186/s40635-018-0184-3

Mei, J., Riedel, N., Grittner, U. et al. (2018). Body temperature measurement in mice during acute illness: Implantable temperature transponder versus surface infrared thermometry. *Sci Rep 8*, 3526. doi:10.1038/s41598-018-22020-6

Mei, J., Banneke, S., Lips, J. et al. (2019). Refining humane endpoints in mouse models of disease by systematic review and machine learning-based endpoint definition. *ALTEX 36*, 555-571. doi:10.14573/altex.1812231

Moons, K. G. M., Altman, D. G., Reitsma, J. B. et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med 162*, W1-73. doi:10.7326/M14-0698

Morton, D. and Griffiths, P. (1985). Guidelines on the recognition of pain, distress and discomfort in experimental animals and an hypothesis for assessment. *Vet Rec 116*, 431-436. doi:10.1136/vr.116.16.431

Morton, D. B. (2000). A systematic approach for establishing humane endpoints. *ILAR J 41*, 80-86. doi:10.1093/ilar.41.2.80

Napier, B. A., Brubaker, S. W., Sweeney, T. E. et al. (2016). Complement pathway amplifies caspase-11-dependent cell death and endotoxin-induced sepsis severity. *J Exp Med 213*, 2365-2382. doi:10.1084/jem.20160027

Nemzek, J. A., Xiao, H.-Y., Minard, A. E. et al. (2004). Humane endpoints in shock research. *Shock 21*, 17-25. doi:10.1097/01.shk.0000101667.49265.fd

Nunamaker, E. A., Artwohl, J. E., Anderson, R. J. et al. (2013). Endpoint refinement for total body irradiation of C57BL/6 mice. *Comp Med 63*, 22-28.

Paster, E. V., Villines, K. A. and Hickman, D. L. (2009). Endpoints for mouse abdominal tumor models: Refinement of current criteria. *Comp Med 59*, 234-241.

Pfeiffenberger, U., Yau, T., Fink, D. et al. (2015). Assessment and refinement of intra-bone marrow transplantation in mice. *Lab Anim 49*, 121-131. doi:10.1177/0023677214559627

Ray, M. A., Johnston, N. A., Verhulst, S. et al. (2010). Identification of markers for imminent death in mice used in longevity and aging research. *J Am Assoc Lab Anim Sci 49*, 282-288.

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W. et al. (2008). Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J 50*, 419-430. doi:10.1002/bimj.200710415

Schmid, M., Wright, M. and Ziegler, A. (2016). On the use of Harrell's C for clinical risk prediction via random survival forest. http://arxiv.org/pdf/1507.03092.pdf (accessed November 2019)

Schröder, M. S., Culhane, A. C., Quackenbush, J. et al. (2011). survcomp: An R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics 27*, 3206-3208. doi:10.1093/bioinformatics/btr511

Sliepen, S. H. J., Diaz-Delcastillo, M., Korioth, J. et al. (2019). Cancer-induced bone pain impairs burrowing behaviour in mouse and rat. *In Vivo 33*, 1125-1132. doi:10.21873/invivo.11582

Smith, D., Anderson, D., Degryse, A.-D. et al. (2018). Classification and reporting of severity experienced by animals used in scientific procedures: FELASA/ECLAM/ESLAV working group report. *Lab Anim 52*, 5-57. doi:10.1177/0023677217744587

Tag, C. G., Sauer-Lehnen, S., Weiskirchen, S. et al. (2015a). Bile duct ligation in mice: Induction of inflammatory liver injury and fibrosis by obstructive cholestasis. *J Vis Exp*, e52438. doi:10.3791/52438

Tag, C. G., Weiskirchen, S., Hittatiya, K. et al. (2015b). Induction of experimental obstructive cholestasis in mice. *Lab Anim 49*, 70-80. doi:10.1177/0023677214567748

Takayama-Ito, M., Lim, C.-K., Nakamichi, K. et al. (2017). Reduction of animal suffering in rabies vaccine potency testing by introduction of humane endpoints. *Biologicals 46*, 38-45. doi:10.1016/j.biologicals.2016.12.007

Therneau, T. (2019). Package 'survial'. https://cran.r-project.org/web/packages/survival/survival.pdf (accessed August 2019)

Trammell, R. A. and Toth, L. A. (2011). Markers for predicting death as an outcome for mice used in infectious disease research. *Comp Med 61*, 492-498.

Van Belle, V., Pelckmans, K., van Huffel, S. et al. (2011). Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics 27*, 87-94. doi:10.1093/bioinformatics/btq617

Warn, P. A., Brampton, M. W., Sharp, A. et al. (2003). Infrared body temperature measurement of mice as an early predictor of death in experimental fungal infections. *Lab Anim 37*, 126-131. doi:10.1258/00236770360563769

Whittaker, A. L., Lymn, K. A., Nicholson, A. et al. (2015). The assessment of general well-being using spontaneous burrowing behaviour in a short-term model of chemotherapy-induced mucositis in the rat. *Lab Anim 49*, 30-39. doi: 10.1177/0023677214546913

Zhang, H., Xia, W., Lu, X. et al. (2013). A novel statistical prognostic score model that includes serum CXCL5 levels and clinical classification predicts risk of disease progression and survival of nasopharyngeal carcinoma patients. *PLoS One 8*, e57830. doi:10.1371/journal.pone.0057830

## Conflict of interest

The authors declare that they have no conflicts of interest.

## Acknowledgements

This study was supported by the Deutsche Forschungsgemeinschaft (DFG research group FOR 2591, project number: 321137804, ZE 712/1-1 and VO 450/15-1). Xianbin Zhang and Guanglin Tang were supported by the China Scholarship Council (grant numbers: 201608080159 and 201808080167).