Research Article

# Refining Humane Endpoints in Mouse Models of Disease by Systematic Review and Machine Learning-Based Endpoint Definition

*Jie Mei*[1], *Stefanie Banneke*[2], *Janet Lips*[1,3,4], *Melanie T. C. Kuffner*[1,4,5], *Christian J. Hoffmann*[1,4,6], *Ulrich Dirnagl*[1,3,4,7,8], *Matthias Endres*[1,4,6,7,8], *Christoph Harms*[1,3,4,6] *and Julius V. Emmrich*[1,2,6]

[1]Department of Neurology and Department of Experimental Neurology, Neurocure Cluster of Excellence, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany; [2]German Federal Institute for Risk Assessment, German Center for the Protection of Laboratory Animals (Bf3R), Berlin, Germany; [3]QUEST – Center for Transforming Biomedical Research, Berlin Institute of Health (BIH); [4]Center for Stroke Research, Charité – Universitätsmedizin Berlin, Berlin, Germany; [5]Berlin-Brandenburg School for Regenerative Therapies (BSRT), Berlin, Germany; [6]Berlin Institute of Health (BIH), Berlin, Germany; [7]German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany; [8]German Center for Cardiovascular Research (DZHK), Berlin, Germany

**Abstract**

Ideally, humane endpoints allow early termination of experiments by minimizing an animal's discomfort, distress and pain while ensuring that scientific objectives are reached. Yet, lack of commonly agreed methodology and heterogeneity of cut-off values published in the literature remain a challenge to the accurate determination and application of humane endpoints.

With the aim to synthesize and appraise existing humane endpoint definitions for commonly used physiological parameters, we conducted a systematic review of mouse studies of acute and chronic disease models that used body weight, temperature and/or sickness scores for endpoint definition. We searched for studies in two electronic databases (MEDLINE/Pubmed and Embase). Out of 110 retrieved full-text manuscripts, 34 studies were included. We found large intra- and inter-model variance in humane endpoint determination and application due to varying animal models, lack of standardized experimental protocols, and heterogeneity of performance metrics (part 1).

We then used previously published and unpublished data on weight, temperature, and sickness scores from mouse models of sepsis and stroke and applied machine learning models to assess the usefulness of this method for parameter selection and endpoint definition across models. Machine learning models trained with physiological data and sickness severity score or modified DeSimoni neuroscore identified animals with a high risk of death at an early time point in both mouse models of stroke (male: 93.2% at 72 h post-treatment; female: 93.0% at 48 h post-treatment) and sepsis (96.2% at 24 h post-treatment), thus demonstrating generalizability of endpoint determination across models (part 2).

## 1 Introduction

In experimental mouse studies an important challenge for researchers is to identify an endpoint by which the experiment shall be terminated in order to minimize unnecessary suffering of animals without compromising the quality of the experimental data. To systematically address this challenge, the concept of humane endpoints was introduced almost 20 years ago in Europe (OECD, 2000). The application of humane endpoints describes the use of clear, predictable, and irreversible criteria, which can be used as a substitute for a more severe experimental outcome such as extreme suffering or death. Systematic implementation of humane endpoints can prevent or reduce pain and/or suffering whilst still meeting experimental objectives (Nemzek et al., 2004).

Thus, application of humane endpoints is a key component of refining studies to comply with 3R principles. In models of acute disease, death may occur within hours following an experimental intervention, which requires both intensive follow-up and consis-

Received December 23, 2018; Accepted April 10, 2019;
Epub April 18, 2019; © The Authors, 2019.

**ALTEX 36(4), 555-571.** doi:10.14573/altex.1812231

Correspondence: Julius Emmrich, Charité Universitätsmedizin Berlin,
Department of Neurology and Department of Experimental Neurology,
Charitéplatz 1, 10117 Berlin, Germany
(julius.emmrich@charite.de)
*or* Christoph Harms, Center for Stroke Research, Department of
Experimental Neurology, Charitéplatz 1, 10117 Berlin, Germany
(christoph.harms@charite.de)

tency in endpoint determination. However, the varying nature of animal models and disease progression, lack of reporting these details in the literature, lack of standardized evaluation protocols, and heterogeneity of endpoints published in the literature make it difficult to accurately determine and apply humane endpoints (Franco et al., 2012).

So far, various approaches to humane endpoint evaluation have been proposed. These are based on physiological parameters such as body weight, temperature, or standardized sickness scores. Most commonly, analysis is conducted in a non-comprehensive manner, e.g., by arbitrary selection of a parameter and a cut-off value corresponding to the highest mortality rate or best separation between treated and sham-treated animals. However, these approaches often require manual, time-consuming computation and are prone to inter-observer bias. Machine learning, a technique used to identify underlying patterns from given datasets to produce reliable, repeatable predictions, has been applied in a number of different animal studies to classify individual/social behaviors (Kabra et al., 2013), automatize behavior analysis (Han et al., 2018), or to identify behavioral strategies and decision-making processes (Yamaguchi et al., 2018). To our knowledge, using machine learning methods for humane endpoint characterization has not yet been systematically assessed.

The aim of the present study therefore was twofold: first, to identify and appraise existing humane endpoint definitions in mouse models of acute and chronic disease by conducting a systematic review of studies using weight, body temperature, and/or sickness scores for humane endpoint refinement and evaluation, and second, to examine the potential usefulness and accuracy of using machine learning with an automated parameter search to automatically define humane endpoints. To maximize generalizability of results, we used previously published and unpublished data from two independent mouse models of acute disease, namely, a middle cerebral artery occlusion (MCAo) stroke model and a lipopolysaccharide (LPS)-induced systemic inflammation model, respectively (Donath et al., 2016; Mei et al., 2018).

We found great heterogeneity of published cut-off criteria and thresholds, illustrating a distinct difficulty in adopting humane endpoints from the literature. However, we show that machine learning can be used to accurately determine humane endpoint criteria and cut-off threshold values at early time points following stroke or systemic inflammation thus potentially reducing otherwise unnecessary suffering.

## 2 Animals, materials, and methods

### 2.1 Systematic review

#### 2.1.1 Search strategy

Studies were identified, screened and extracted for relevant data following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (http://www.prisma-statement.org). Literature search, title and abstract screening was conducted by JM. Full text screening was conducted by JM and JVE. A search was conducted on the MEDLINE/PubMed databases for all research articles from 1946 to Feb 07, 2018 using the following Boolean string with Medical Subject Headings (MeSH): ((“Mice”[Mesh]) AND (“Endpoint Determination”[Mesh] OR “Animal Use Alternatives”[Mesh] OR humane endpoint* OR humane end point* OR surrogate endpoint* OR surrogate end point* OR thermometry OR thermometer OR telemetry OR refinement OR welfare) AND (“Body Weight”[Mesh] OR “Body Temperature”[Mesh] OR body temperature OR weight NOT fetal NOT fetus OR score* OR scoring)), and on the Embase database for all research articles from 1947 to Feb 07, 2018 using EMBASE Thesaurus (EMTREE) with Boolean string: (exp mice) and (exp Body Temperature or exp Body Weight or (score$ or scoring) or body temperature or (weight not fetal not fetus)) and (humane end point$ or humane endpoint$ or surrogate end point$ or surrogate endpoint$ or (thermometry or thermometer or telemetry) or (welfare or refinement)).

#### 2.1.2 Exclusion and inclusion criteria

Studies that fulfilled the following inclusion criteria were included in the systematic review: (a) original research articles on mouse models of acute and/or chronic disease, (b) physiological parameters such as body temperature, body weight, or sickness severity scores were used individually or in combination to identify and/or evaluate humane endpoints, and (c) studies that applied pre-defined humane endpoints determined from body temperature, body weight, or sickness severity scores.

Irrelevant studies were excluded if: (a) subjects used were other than mice, (b) article was a conference abstract, experimental protocol, or review, (c) article was written in a language other than English, (d) parameters used to determine humane endpoints were other than body temperature, body weight, or sickness severity scores, and (e) no humane endpoints were applied in the course of experiments or if the study was unrelated to humane endpoint determination.

#### 2.1.3 Extraction of relevant data

Relevant data was extracted and compared through a data extraction sheet. Extraction procedure was conducted by JM. Extracted data included (a) disease model, (b) sample size, (c) time course of the experiment, (d) frequency of evaluation/measurement, (e) humane endpoint(s) used/proposed, (f) cut-off criteria for euthanasia, (g) metrics for evaluating the humane endpoint(s), and (h) performance of the humane endpoint(s). When multiple endpoints were applied together in one study, all available descriptions were included. Missing data entries were marked with N/A (not applicable).

### 2.2 Animal models of stroke and sepsis

#### 2.2.1 Animals

No animals were used for this study. Rather, all data analyzed in this study were sourced from the authors’ previously published and unpublished results using a middle cerebral artery occlusion (MCAo) stroke model and a lipopolysaccharide (LPS)-induced systemic inflammation model, respectively (Hoffmann et al., 2015; Donath et al., 2016; Koch et al., 2017; Emmrich et al.,

**Tab. 1: Strain and origin of animals used for automated parameter search to define humane endpoints for (A) middle cerebral artery occlusion (MCAo) stroke model, and (B) lipopolysaccharide (LPS)-induced sepsis model**
m, male; f, female

**A**

| Strain | n | Origin |
|---|---|---|
| C57BL/6NCrl | 74 (m: 74) | Charles River Laboratories |
| Tg(Gjb6-cre/ERT2)53-33Fwp [MGI:4420273] x custom-made Tg(ROSA26-FLEX IL6)1Ch | 166 (m: 85; f: 81) | F. Pfrieger; Charité Universitätsmedizin Berlin; Research Institutes for Experimental Medicine |
| C57BL/6N-Zfp580$^{tm1a(EUCOMM)Hmgu}$/BayMmucd | 158 (m: 84; f: 74) | Charité Universitätsmedizin Berlin; Research Institutes for Experimental Medicine |
| Tg(Cdh5-cre/ERT2)1Rha x custom-made Tg(ROSA26-FLEX IL6)1Ch | 33 (m: 16; f: 17) | R. Adams; Charité Universitätsmedizin Berlin; Research Institutes for Experimental Medicine |
| Sorcs2tm1Anyk [MGI:5649357] | 56 (m: 56) | |

**B**

| Strain | n | Origin |
|---|---|---|
| C57BL/6J | 55 (f) | Charles River Laboratories |
| Mertk (*B6;129-Mertk$^{tm1Grl}$/J*) | 126 (f) | The Jackson Laboratory |
| Cd11b (B6;129-Mertk$^{tm1Grl}$/J, B6.129S4-Itgam$^{tm1Myd}$/J) | 126 (f) | Hertie Institute for Clinical Brain Research |
| Mfge8 | 128 (f) | C. Théry, INSERM 932, France |

2017; Mei et al., 2018). For the original studies, all experimental procedures were approved by the ethical review committee of *Landesamt für Gesundheit und Soziales* (LaGeSo), Berlin (Reg G0385/08, G0188/11, G0354/11, G0197/12, G005/16, G0057/16, G0119/16, GG254/16, G0157/17, stroke; Reg G239/15, sepsis) and were conducted in accordance with the German animal protection law and local animal welfare guidelines. Reporting of results based on the authors' own historical data complies with the ARRIVE guidelines (Kilkenny et al., 2010) and with the guidelines for genetically modified organisms (441/06). Data from 922 animals were included in this study. All inspections and measurements were performed in the same facility where animals were housed.

For the stroke model, adult male and female mice were used (total: n = 487; Tab. 1a; Slezak et al., 2007; Skarnes et al., 2011; Benedito et al., 2009; Glerup et al., 2014). Seven mice were not assigned to any treatment group (male: 5; female: 2) as they died of natural causes or reached a humane endpoint prior to the start of the experiment. Therefore, 480 out of 487 mice were randomly assigned to a 30 min MCAo (n = 73; male: 53; female: 20), a 45 min MCAo (n = 331; male: 213, female: 118) or a sham procedure (n = 76; male: 44; female: 32), at the age of 8-12 weeks. Mice were housed in groups of up to 12 animals per cage at 22 ±2°C, humidity of 55 ±10%, and a 12-hour light/dark cycle (12:12 h, lights on: 7:00 h, lights off: 19:00 h). Aspen woodchips were used as bedding.

For the sepsis model, female homozygous knockout mice and their homozygous wildtype littermates were used in experiments at the age of 8-10 weeks (total: n = 435; Tab. 1b). Mice were housed in groups of up to 12 animals per cage at 23 ± 1°C, humidity of 60 ± 5%, and a 12-hour light/dark cycle (12:12 h light/dark cycle, lights on: 20:00, lights off: 8:00) and were exposed to white noise at moderate intensity (65dB) during the dark phase (Dohm Sleepmate, Marpac Sound Machines, Wilmington, USA). During acute illness and recovery, mice were housed individually. Wood shavings were used as bedding.

### 2.2.2 Treatments

Stroke model: Mice were subjected to 30 or 45 minutes temporary filamentous middle cerebral artery occlusion (MCAo) or sham procedure. The filamentous MCAo model was performed as described in Dirnagl et al. (2012). For sham animals, the filament was advanced to the MCA and withdrawn immediately.

Sepsis model: Lipopolysaccharide (LPS) or physiological phosphate-buffered saline solution (PBS) were administered intraperitoneally for the induction of a systemic inflammatory response or control, respectively. The injection was performed as previously described (Mei et al., 2018).

Animals were randomized to treatment groups using the GraphPad calculator tool[1] or Research Randomizer tool[2] for the stroke and sepsis model, respectively. To minimize experimenter bias, randomization was conducted by a researcher who was not

[1] http://www.graphpad.com/quickcalcs/randomize1.cfm

[2] https://www.randomizer.org

involved in injections, treatments, data acquisition or analysis. Information on strain, genotype and treatment group assignment was concealed from experimenters until the end of the study.

### 2.2.3 Physiological parameters and scoring

In the stroke model, body weight and a modified version of the DeSimoni neuroscore, a composite score of general behavioral alterations and focal motor, sensory, reflex, and balance deficits to evaluate neurological outcome following cerebral ischemia in mice, were obtained as previously described (Donath et al., 2016). Core body temperature was quantified non-invasively using subcutaneous radio-frequency identification (RFID) transponders as described (Donath et al., 2016).

In the sepsis model, a sickness score adapted from the murine sepsis score was obtained based on general activity and response to stimuli as previously described (Mei et al., 2018). Surface body temperature was quantified using two non-contact infrared thermometer models as described previously (Mei et al., 2018). For body weight acquisition, a bench scale (PCB 1000-1, KERN & SOHN GmbH, Balingen, Germany) was used. Animals were weighed once their body and tail were in a plastic box placed on the top of the scale.

In both disease models, the duration of manual handling was minimized to reduce stress and discomfort when examining signs of sickness of experimental animals. Low anxiety handling methods including cupping the animal between both hands and using a handling box were applied. In addition, and only if necessary, animals were lifted by the base of the tail for no longer than 2-3 seconds.

### 2.2.4 Timeline of physiological monitoring

In the stroke model, baseline body weight and temperature were measured at 7:00-8:00 on the day of MCAo, followed by 2 inspections on the day of surgery and consequent daily inspection for qualitative humane endpoint criteria at 7:00-9:00 until day 28. The modified DeSimoni score for individual animals was assessed on the day of MCAo and on the 1st, 2nd, 7th, 14th, and 21st day post-MCAo as previously described (Donath et al., 2016).

In the sepsis model, baseline temperature and weight were measured at 8:00 on the day of the first injection. Body temperature and sickness score were obtained eight times daily (8:00 to 20:00, every 90 min) on the two consecutive injection days, then three times daily (8:00 to 20:00, every 6 h) for two days after the second injection, and once a day (8:00) from post-injection day 3 until day 30 after the second injection. Body weight was obtained three times daily (8:00 to 20:00, every 6 h) during the two injection days and the first two days following the second injection, then once per day at 8:00 until day 30 after the second injection (Mei et al., 2018). To avoid stress-induced fluctuations in body temperature, animals were weighed after body temperature acquisition.

Body temperature, body weight, and sickness score were assessed for 21 or 30 days post-MCAo or LPS/PBS injection, respectively. In accordance with the aim of the study, data from time points later than that of the death of the last animal in each experiment was not included.

### 2.2.5 Humane endpoint criteria

In the stroke model, animals were euthanized by cervical dislocation upon reaching a score of 2 of the 2nd criteria, or a score of 3 or 4 of the 3rd-12th criteria in the modified DeSimoni neuroscore (Donath et al., 2016). In addition to the score-based criteria, animals were euthanized when a loss of more than 20% baseline body weight occurred or the following qualitative humane endpoint criteria were observed during inspection: complete paralysis with absence of spontaneous movement, severe ataxia or loss of postural reflexes, severe epileptic seizures, severe reduction of general health status with reduced grooming or refusal of food intake.

In the sepsis model, upon reaching a sickness score greater than 4 once or a score of 4 twice within 2 hours, animals were immediately removed from the cage and euthanized by cervical dislocation (Mei et al., 2018).

### 2.2.6 Exclusion criteria

In the stroke model, animals that were (a) attacked by littermates before or during the experiment (n = 8); (b) failed to learn the behavioral task prior to MCAo (n = 9); (c) died during or within the first hour after anesthesia as a result of surgical complications (n = 12); (d) euthanized on the day of surgery (n = 1); (e) euthanized after the 30th day post-MCAo (n = 7); and (f) of a baseline temperature < 32°C (n = 4), were not included in subsequent analysis, leading to exclusion of 41 out of 487 animals (8.4%).

In the sepsis model, no animal was excluded.

### 2.2.7 Data analysis and statistics

Results are expressed as mean (SD) unless otherwise specified. Data processing and statistical analysis was performed using SPSS version 24 (SPSS Inc., Chicago, IL, USA) and Python 2.7.10 (Python Software Foundation, Beaverton, OR, USA). Risk of death as an outcome event was evaluated with the scikit-learn toolkit (sklearn; Pedregosa et al., 2011) for physiological parameters including core body temperature, surface body temperature, body weight and modified DeSimoni neuroscore or sickness severity score.

A primary aim of this study was to identify physiological parameters that can be used to separate animals that are at a higher risk of death from animals that would reach the planned experimental endpoint. Therefore, apart from assessing the prediction accuracy of various models, we also identified the predictive power of physiological parameters, individually or in combination. To assess and identify (a) general performance of machine learning models, (b) usability of physiological parameters obtained at different time points in death prediction, and (c) model hyperparameters, grid search with stratified 3-fold cross-validation was applied. Core temperature (stroke model), surface temperature (sepsis model), body weight (both models), sickness score (sepsis model) or modified DeSimoni neuroscore (stroke model), and the absolute change of these parameters per timepoint (calculated by subtracting the baseline value from the measured value at a given timepoint) were used individually or in combination to train machine learning models. Models used for this study included logistic regression, Gaussian Naïve Bayes, decision tree (of max_depth = 1, 2, 3, or 4), support vector machine (with linear or radial basis function (RBF) kernels; C = 1,

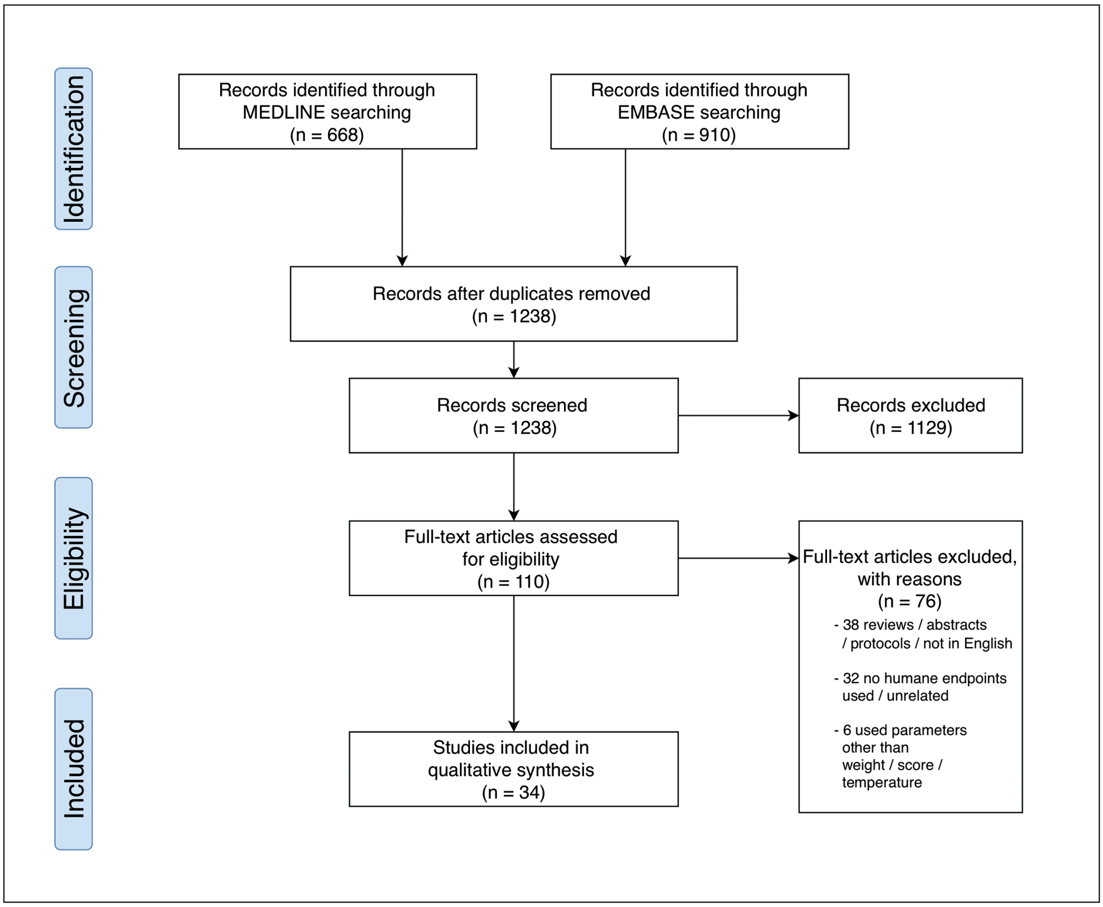
Identification
Screening
Eligibility
Included
Records identified through MEDLINE searching (n = 668)
Records identified through EMBASE searching (n = 910)
Records after duplicates removed (n = 1238)
Records screened (n = 1238)
Records excluded (n = 1129)
Full-text articles assessed for eligibility (n = 110)
Full-text articles excluded, with reasons (n = 76)
- 38 reviews / abstracts / protocols / not in English
- 32 no humane endpoints used / unrelated
- 6 used parameters other than weight / score / temperature
Studies included in qualitative synthesis (n = 34)

**Fig. 1: Flow diagram showing the number of studies identified, screened, extracted and included in this systematic review**

10, or 100; gamma = 0.01, 0.001, or 0.0001), and random forest classifier (with n_estimators = 2, 4, or 8).

Available data from all time points before the average time of death of non-survivor animals was included in the analysis. First, to reduce complexity of the analysis and enhance the applicability of the method, measurements obtained at the same time point were used to train the predictive models. For example, temperature readings obtained at 24 hours after stroke/sepsis could be combined with sickness scores obtained at the same time point, but not sickness scores obtained at other time points. Second, an expanded parameter search with combinations of physiological parameters obtained at different time points was conducted. When training support vector machines, input features were scaled to a zero mean and unit variance.

#### 2.2.8 Data availability

Two datasets including (1) core body temperature, body weight, and modified DeSimoni neuroscore of animals of the stroke model[3] and (2) surface body temperature, body weight and sickness severity score of animals of the sepsis model[4] are available as open data on Figshare Repository in raw data format.

## 3 Results

### 3.1 Systematic review

1,578 search results were retrieved (Medline: 668; Embase: 910) and 1,238 were included in title and abstract screening after duplicates were removed. Overall, 110 full text articles were screened and a total of 34 studies were included for subsequent data extraction (Fig. 1; Tab. 2). Included studies represented a wide range of acute and chronic mouse models including infection/inflammation (n = 14; Nemzek et al., 2004; Huet et al., 2013; Bast et al., 2004; Adamson et al., 2013; Kort et al., 1998; Hankenson et al., 2013; Warn et al., 2003; Arranz-Solis et al., 2015; Dellavalle et al., 2014; Molins et al., 2012; Wright and Phillpotts, 1998; Sand et al., 2015; Miller et al., 2013; Trammell and Toth, 2011), toxin/poisoning (n = 3; Vlach et al., 2000; Beyer et al., 2009; Cates et al., 2014), cancer/tumor (n = 5; Husmann et al., 2015; Aldred et al., 2002; Miller et al., 2016; Paster et al., 2009; Hunter et al., 2014), and others (n = 12; Solomon et al., 2011; Stoica et al., 2016; Leon et al., 2005; Takayama-Ito et al., 2017; Passman et al., 2015; Chappell et al., 2011, Faller et al., 2015; Weismann et al., 2015; Koch et al., 2016; Nunamaker et al., 2013a,b; Ray et al., 2010).

[3] doi:10.6084/m9.figshare.7479965
[4] doi:10.6084/m9.figshare.7480016

**Tab. 2: List of studies included in the review (n = 34) summarized by type of experiment, sample size, time course of the experiment, type of humane endpoints, frequency of inspection/measurement, and cut-off threshold used/proposed for euthanasia**
N/A, not found/not available; sur, survived

| Experiment | Sample size | Time course | Type of humane endpoint | Frequency of measurement | Cut-off threshold | Reference |
|---|---|---|---|---|---|---|
| Intranasal invasive pulmonary aspergillosis | n = 122; n(sur) = 45 | < 8 days | weight, surface temperature | once daily (weight); ≤ 3 times/day (surface temperature) | > 20% weight loss; surface temperature < 28.8°C | Adamson et al., 2013 |
| Leukemia | n = 20 | ≤ 14 days | score, clinical signs | every 12 h (first 2 days); every 6 h (day 3 and thereafter) | score ≤ 3; clinical signs on two consecutive examinations | Aldred et al., 2002 |
| *Neospora caninum* infection | n = 118; n(sur) = 93 | ≤ 30 days | score | twice daily | score ≥ 3 | Arranz-Solis et al., 2015 |
| Pneumonia | n = 31; n(sur) = 10 | ≤ 96 hours | surface temperature | twice daily | surface temperature ≤ 30°C | Bast et al., 2004 |
| Ricin poisoning | n = 66 | ≤ 100 hours | core temperature, clinical signs | every 30 min | two consecutive temperature measurements < 32°C; clinical signs | Beyer et al., 2009 |
| Rattlesnake venom | n = 30; n(sur) = 19 | ≤ 8 hours | core temperature | every 10-30 min (first 2 h post-injection); every 1-2 h (thereafter) | core temperature < 33.2°C | Cates et al., 2014 |
| Postsurgical recovery | n = 45 | ≤ 14 days | score | daily | score ≥ 4 | Chappell et al., 2011 |
| Plasmodium infection | n = 40; n(sur) = 27 | ≤ 15 days | surface temperature | once daily (until symptoms were present); 3 times/day (thereafter) | surface temperature < 30°C | Dellavalle et al., 2014 |
| Myocardial infarction | n = 60 | ≤ 8 weeks | weight, clinical signs | at 24 h and 30 min after application of analgesia | weight loss (unspecified); clinical signs | Faller et al., 2015 |
| Ocular herpes simplex virus infection | n = 120; n(sur) = 38 | ≤ 60 days | core temperature, weight, score | once daily (until day 4; day 15-30); twice daily (day 5 -day 15) | core temperature < 34.5°C; > 0.05g/day weight loss, combination of temperature and weight loss; score = 3 for 24 h | Hankenson et al., 2013 |
| Pneumonia (septic shock) | n = 118; n(sur) = 104 | ≤ 5 days | score | twice daily (score 1); 3 times/day (score 2); 4-6 times/day (score 3); hourly (score 4) | score = 4 | Huet et al., 2013 |
| Lymphoma | n = 36 | ≤ 5 weeks | weight, clinical signs | 5-7 times/week | > 20% weight loss or > 15% weight loss for 72 h; clinical signs | Hunter et al., 2014 |
| Bone cancer | n = 30 | ≤ 26 or 34 days | weight, clinical signs | once a week; twice daily after application of analgesia | > 15% weight loss; clinical signs | Husmann et al., 2015 |
| Total-body irradiation | n = 132; n(sur) = 77 | ≤ 30 days | score | twice daily (noncritical period); ≤ 4 times/day (critical period) | score = 12 | Koch et al., 2016 |

| Experiment | Sample size | Time course | Type of humane endpoint | Frequency of measurement | Cut-off threshold | Reference |
|---|---|---|---|---|---|---|
| Pneumonia | n = 10 | ≤ 24 hours | core temperature | twice daily | core temperature < 36°C | Kort et al., 1998 |
| Heat stress | n = 78 | < 44 hours | core temperature | continuous (every 60 sec) | no recovery from hypothermia by 765 min | Leon et al., 2005 |
| Influenza A infection | n = 16 | ≤ 7 days | weight, score | daily | > 25% weight loss; score ≥ 4 | Miller et al., 2013 |
| Bladder cancer | n = 80 | ≤ 50 days | size of tumor, weight, clinical signs | daily | tumor > 10 mm (12mm), > 15% weight loss, or if either coincided with clinical signs | Miller et al., 2016 |
| *Francisella tularensis* infection | n = 56 | < 264 hours | core temperature | every 1 to 2 hours | N/A | Molins et al., 2012 |
| Septic shock | n = 36; n(sur) = 10 | ≤ 14 days | weight, surface temperature | once daily | surface temperature ≤ 30°C; initial weight gain | Nemzek et al., 2004 |
| Total-body irradiation | n = 240; n(sur) = 57 | ≤ 30 days | score | once daily (days 1-6 and 23-30); twice daily (days 7-22) | score ≥ 7 | Nunamaker et al., 2013a |
| Total-body irradiation | n = 175; n(sur) = 66 | ≤ 30 days | score | once daily (days 1-6; 19-30; twice daily (days 7-18) | score > 6 | Nunamaker et al., 2013b |
| Choline-deficient, ethionine-supplemented diet | n = 34; n(sur) = 23 | < 3 weeks | weight, score | twice daily until partial recovery; daily thereafter | ≥ 20% weight loss; score = 3 | Passman et al., 2015 |
| Abdominal tumor | n = 40 | ≤ 46 days | score, clinical signs | daily | score = 1 with clinical signs; score = 3 | Paster et al., 2009 |
| Longevity and aging | n = 110 | ≤ 40 months | core temperature, weight, temperature x weight | at least once every 4 weeks | ≥ 15% weight loss; core temperature < 25°C; | Ray et al., 2010 |
| Septic shock | n = 15 | ≤ 21 days | score | once after 24 h; irregular monitoring (in between) | score = 5 | Sand et al., 2015 |
| Amyotrophic lateral sclerosis | n = 162 | ≤ 150 days of age | score | N/A | score = 4 | Solomon et al., 2011 |
| Amyotrophic lateral sclerosis | n = 42 | ≤ 260 days of age | weight, clinical signs | once every week | ≥ 15% weight loss; clinical signs | Stoica et al., 2016 |
| Rabies virus infection | n = 359 | ≤ 11 days | score, weight | daily | score = 2 combined with a weight loss of ≥ 15% | Takayama-Ito et al., 2017 |
| Influenza virus infection | n = 118 | ≤ 5-21 days after injection, depending on the type of infection | core temperature, weight, core temperature x weight (T x BW) | daily (temperature); weight (3 times/ week) | temperature < 35°C or T x BW < 60% of baseline values on day 7 after infection;<br>T x BW < 90% of baseline value on day 2 or day 5 after infection;<br>T x BW < 85% of baseline value on day 1 after infection | Trammell and Toth, 2011 |

| Experiment | Sample size | Time course | Type of humane endpoint | Frequency of measurement | Cut-off threshold | Reference |
|---|---|---|---|---|---|---|
| Septic shock | n = 48;<br>n(sur) = 22 | ≤ 25 days | core temperature | every 15 minutes | core temperature < 23.4°C | Vlach et al., 2000 |
| Fungal infection | n = 160;<br>n(sur) = 100 | ≤ 11 days | core temperature | ≤ 4 times daily, maximum of 10 h between observations | core temperature < 33.3°C | Warn et al., 2003 |
| GM1 gangliosidosis | n = 122 | ≤ 576 days | weight, clinical signs | 19 times at irregular intervals | ≥ 15% weight loss from maximum weight; clinical signs | Weismann et al., 2015 |
| Venezuelan encephalomyelitis virus infection | n = 20 | ≤ 14 days | score | at least twice daily | score = 2 | Wright and Phillpotts, 1998 |

### 3.1.1 Time course of the study and frequency of monitoring

Time courses of experiments ranged from 8 hours to 40 months. Among the 34 studies, duration of experiments was shorter than or equal to 24 hours in 2 studies (Kort et al., 1998; Cates et al., 2014), between a day and a week in 5 studies (Leon et al., 2005; Beyer et al., 2009; Miller et al., 2013; Bast et al., 2004; Huet et al., 2013), between a week and one month in 17 studies (Molins et al., 2012; Takayama-Ito et al., 2017; Passman et al., 2015; Wright and Phillpotts, 1998; Sand et al., 2015; Chappell et al., 2011; Adamson et al., 2013; Koch et al., 2016; Trammell and Toth, 2011; Nunamaker et al., 2013a,b; Nemzek et al., 2004; Dellavalle et al., 2014; Aldred et al., 2002; Warn et al., 2003; Vlach et al., 2000; Arranz-Solis et al., 2015), between one month and 3 months in 6 studies (Miller et al., 2016; Faller et al., 2015; Hankenson et al., 2013; Paster et al., 2009; Husmann et al., 2015; Hunter et al., 2014), and longer than 3 months in 4 studies (Weismann et al., 2015; Solomon et al., 2011; Stoica et al., 2016; Ray et al., 2010).

A great variance was observed in the frequency of evaluation intervals, with the most frequent data collection occurring once per minute using an automated system (Leon et al., 2005), while the longest interval between inspections was once or at least once every 4 weeks (Ray et al., 2010; Weismann et al., 2015). Some authors used an inspection frequency of once per day (Miller, D. S. et al., 2013; Miller, A. et al., 2016; Chappell et al., 2011; Paster et al., 2009; Nemzek et al., 2004), while others used varying interval frequencies such as once per 15 minutes (Vlach et al., 2000), once per 30 minutes (Beyer et al., 2009), twice per day (Cates et al., 2014; Kort et al., 1998; Arranz-Solis et al., 2015), at least twice per day (Wright and Phillpotts, 1998), at least 4 times per day (Warn et al., 2003), once per week (Stoica et al., 2016), 5-7 times per week (Hunter et al., 2014). As expected, in studies with faster disease progression, authors adjusted the evaluation schedule accordingly by increasing the frequency of monitoring for animals in severe distress or animals requiring additional care (Molins et al., 2012; Passman et al., 2015; Hankenson et al., 2013; Huet et al., 2013; Koch et al., 2016; Trammell and Toth, 2011; Cates et al., 2014; Nunamaker et al., 2013a,b; Dellavalle et al., 2014; Aldred et al., 2002; Husmann et al., 2015). In two studies (Adamson et al., 2013; Trammell and Toth, 2011), individual physiological parameters were assessed at different time points. In three studies, animals were inspected only once after treatment (Takayama-Ito et al., 2017; Sand et al., 2015; Faller et al., 2015).

### 3.1.2 Body temperature

Among the 34 studies, 14 demonstrated that both core body temperature (n = 10; Molins et al., 2012; Leon et al., 2005; Beyer et al., 2009; Cates et al., 2014; Hankenson et al., 2013; Trammell and Toth, 2011; Warn et al., 2003; Kort et al., 1998; Vlach et al., 2000; Ray et al., 2010) and surface body temperature (n = 4; Adamson et al., 2013; Bast et al., 2004; Nemzek et al., 2004; Dellavalle et al., 2014) could be used to refine the humane endpoint. In 6 of the 14 studies, other physiological parameters were used in combination or independently as humane endpoint criteria, including clinical signs (Beyer et al., 2009), weight (Hankenson et al., 2013; Adamson et al., 2013; Nemzek et al., 2004), and the product of body temperature and weight (Trammell and Toth, 2011; Ray et al., 2010). Cut-off values for euthanasia ranged from 23.4 to 36°C (31.55 (4.7) °C) for endpoints determined from core temperature and from 28.8 to 30°C (29.7 (0.6) °C) for endpoints determined from surface temperature. In addition, recovery from hypothermia (Leon et al., 2005) or a temperature drop below the baseline mean temperature (Molins et al., 2012) were used as humane endpoints. While in most studies, animals were humanely killed upon reaching the cut-off criterion at one single time point, Beyer et al. (2009) euthanized animals when the body temperature was lower than 32°C in two consecutive inspections.

### 3.1.3 Body weight

In 14 out of the 34 studies, body weight was used to determine the humane endpoint (Takayama-Ito et al., 2017; Passman et al., 2015; Miller, D. S. et al., 2013; Miller, A. et al., 2016; Faller et al., 2015; Weismann et al., 2015; Adamson et al., 2013; Hankenson et al., 2013; Trammell and Toth, 2011; Nemzek et al., 2004; Stoica et al., 2016; Husmann et al., 2015; Ray et al., 2010; Hunter et al., 2014). All of these 14 studies included additional humane endpoints such as clinical signs of distress and disease progression (Miller et al., 2016; Faller et al., 2015; Weismann et al., 2015; Stoica et al., 2016; Husmann et al., 2015; Hunter et al., 2014), sickness severity scores (Passman et al., 2015; Miller et al., 2013; Takayama-Ito et al., 2017), body temperature (Hankenson et al., 2013; Adamson et al., 2013; Nemzek et al., 2004), and the product of body temperature and weight (Trammell and Toth, 2011; Ray et al., 2010). Although a weight loss of more than 20% compared to baseline is widely regarded as a common humane endpoint, it was only reported in 3 studies (Passman et al., 2015; Adamson et al., 2013; Ray et al., 2010). Other authors used a weight loss of more than 15% (Takayama-Ito et al., 2017; Miller et al., 2016; Weismann et al., 2015; Stoica et al., 2016; Husmann et al., 2015) or 25% (Miller et al., 2013). One study used an absolute weight loss of greater than 0.05 g per day as an indicator of a higher risk of death (Hankenson et al., 2013). In a mouse model of cecal ligation and puncture (CLP), initial weight gain was observed in 100% animals that died within the next 3 days and was therefore considered as an indicator of higher risk of death (Nemzek et al., 2004). One study did not define a cut-off threshold for a weight-based humane endpoint (Faller et al., 2015). Another used the product of body temperature and weight for humane endpoint definition (Trammell and Toth, 2011).

### 3.1.4 Sickness severity score

Among 15 studies that used sickness severity scores to determine the humane endpoint (Passman et al., 2015; Wright and Phillpotts, 1998; Sand et al., 2015; Miller et al., 2013; Chappell et al., 2011; Huet et al., 2013; Koch et al., 2016; Nunamaker et al., 2013a,b; Paster et al., 2009; Solomon et al., 2011; Arranz-Solis et al., 2015; Aldred et al., 2002; Hankenson et al., 2013; Takayama-Ito et al., 2017), 3 applied the score-based threshold with other criteria such as weight (Passman et al., 2015; Miller et al., 2013; Takayama-Ito et al., 2017) and clinical signs (Paster et al., 2009). There was great heterogeneity in score-based thresholds, reflecting the common use of model-specific scores. In 12 studies higher scores indicated more severe symptoms. In one study, a score of 0-1 was assigned to animals showing abnormal behavior and appearance, using a total score of 1 as the humane endpoint (Paster et al., 2009). In one study, a score sheet indicating clinical symptoms was used to determine sickness severity, however, the cut-off value for early euthanasia was not clearly described (Aldred et al., 2002).

### 3.1.5 Combining body weight, body temperature, and sickness severity scores in humane endpoint determination

In 16 studies, more than one physiological or behavioral parameter was used in determining the humane endpoint. Thus, the cut-off criterion was defined by fulfilling one or more physiological or behavioral criteria. Among the 15 studies, 3 studies involved a direct combination (e.g., the product of two parameters) of more than one physiological parameter to derive a surrogate indicator for a higher risk of death (Hankenson et al., 2013; Trammell and Toth, 2011; Ray et al., 2010). In studies using a product of more than one parameter, death could be predicted with higher accuracy. For example, the composite obtained by multiplying weight and body temperature yielded higher prediction accuracy than applying weight or body temperature cut-off criteria individually.

### 3.1.6 Evaluation of the humane endpoints

Twenty studies assessed the performance of humane endpoints. Predictability of death as an outcome event was evaluated by means of sensitivity (n = 6, min = 68%, max = 100%, mean (SD) = 89.7 (13.1)%; Adamson et al., 2013; Hankenson et al., 2013; Trammell and Toth, 2011; Dellavalle et al., 2014; Kort et al., 1998; Warn et al., 2003), specificity (n = 3, min = 90.9%, max = 100%, mean (SD) = 96.0 (4.6)%; Adamson et al., 2013; Hankenson et al., 2013; Warn et al., 2003), logistic regression (n = 2, $p < 0.0001$ and $p = 0.0077$; Cates et al., 2014; Vlach et al., 2000), prediction accuracy (n = 2, 92.7% and 2% underestimation; Koch et al., 2016; Ray et al., 2010), percentage/number of mice present with a particular criterion/sign (n = 5, min = 86%, max = 100%, mean (SD) = 95.3% (5.8); Molins et al., 2012; Takayama-Ito et al., 2017; Bast et al., 2004; Solomon et al., 2011; Aldred et al., 2002), relative number of predicted dead animals (n = 1, 96%; Vlach et al., 2000), physiological changes observed in different treatment groups (n = 2, significant difference observed; Paster et al., 2009; Leon et al., 2005), positive predictive value (n = 1, 55.5%; Nemzek et al., 2004), false positive rate (n = 1, 4-33%; Trammell and Toth, 2011), and corresponding mortality rate (n = 2, 86.2-100% and 78.6-100%; Nunamaker et al., 2013a,b) with some studies using multiple evaluation metrics. In one study, specificity was used to assess humane endpoint performance. However, it could not be appropriately assessed as none of the animals reached the pre-defined cut-off criterion (Dellavalle et al., 2014).

## 3.2 Death prediction in animal models of stroke and sepsis

To facilitate direct comparison, animals in both stroke and sepsis models were divided into three groups based on treatment and survival. Group 1 (control, n(stroke) = 66, n(sepsis) = 151) consisted of sham animals or animals treated with saline that reached the planned experimental endpoint, group 2 (survivor group, n(stroke) = 322, n(sepsis) = 254) consisted of animals treated with MCAo or LPS that reached the planned experimental endpoint and group 3 (non-survivor group, n(stroke) = 58, n(sepsis) = 30) consisted of animals that spontaneously died or were euthanized upon reaching the humane endpoint criteria.

### 3.2.1 Body temperature in survivor and non-survivor animals

In the stroke model, the temperature of the survivor and non-survivor groups decreased by an average of 1.1°C and 4.6°C from baseline (35.3 (2.1) °C and 31.6 (6.6) °C, respectively) during the first 5 days following MCAo (Tab. 3). The core body temperature of the control group remained unchanged during the experiment.

In the sepsis model, LPS-treated animals showed a pronounced decrease in surface body temperature during the two consecutive injection days, regardless of survival status. Lowest surface temperature of survivor animals was observed 10.5 hours following

**Tab. 3: Comparison of physiological measures among control, survivor and non-survivor groups**

(A) Middle cerebral artery occlusion (MCAo) stroke model, from the day animals underwent the MCAo procedure up to the 5th day post stroke; (B) lipopolysaccharide (LPS)-induced sepsis model, from the first injection up to 192 h after the first injection. Baseline measurements were conducted before the MCAo/sham treatment or the first LPS/saline injection. Baseline values were measured on the day of treatment. Monitoring period was defined as the 1st to 5th day post-stroke and the 1st to 8th day post-sepsis.

**A**

| | Control (n = 66) | Survivor (n = 322) | Non-survivor (n = 58) |
|---|---|---|---|
| Baseline core temperature (°C)<br>min, max, mean (SD) | 35.7, 38.9<br>37.0 (0.8) | 34.1, 38.9<br>36.4 (1.0) | 28.0, 38.3<br>36.2 (1.9) |
| Core temperature during monitoring (°C)<br>min, max, mean (SD) | 33.7, 38.4<br>36.6 (0.9) | 20.8, 38.4<br>35.7 (1.6) | 17.8, 37.8<br>33.5 (3.7) |
| Baseline body weight (g)<br>min, max, mean (SD) | 16.0, 32.6<br>24.8 (3.2) | 16.7, 34.9<br>25.4 (3.3) | 14.5, 33.9<br>26.9 (3.7) |
| Body weight during monitoring (g)<br>min, max, mean (SD) | 16.1, 30.1<br>22.4 (2.4) | 13.9, 31.0<br>21.4 (3.0) | 14.3, 29.6<br>21.0 (3.1) |
| Baseline Neuroscore<br>min, max, mean (SD) | 0, 3<br>0.9 (1.0) | 0, 4<br>0.33 (0.8) | 0, 2<br>0.4 (0.9) |
| Neuroscore during monitoring<br>min, max, mean (SD) | 0, 17.0<br>4.9 (4.9) | 0, 35<br>9.1 (5.5) | 0, 41<br>15.4 (9.5) |

**B**

| | Control (n = 151) | Survivor (n = 254) | Non-survivor (n = 30) |
|---|---|---|---|
| Baseline surface temperature (°C)<br>min, max, mean (SD) | 27.0, 33.7<br>30.8 (1.1) | 26.3, 33.8<br>30.5 (1.6) | 27.5, 33.3<br>30.6 (1.4) |
| Surface temperature during monitoring (°C)<br>min, max, mean (SD) | 23.9, 36.5<br>30.6 (1.4) | 20.3, 34.6<br>29.1 (2.0) | 19.0, 31.7<br>25.7 (2.7) |
| Baseline body weight (g)<br>min, max, mean (SD) | 17.5, 26.9<br>21.5 (1.7) | 14.9, 28.7<br>21.5 (1.8) | 17.6, 24.3<br>21.4 (2.0) |
| Body weight during monitoring (g)<br>min, max, mean (SD) | 16.4, 27.6<br>21.9 (1.5) | 13.2, 27.7<br>19.7 (2.1) | 12.5, 23.3<br>18.8 (2.6) |
| Sickness score during monitoring<br>min, max, mean (SD) | 0, 2<br>0 (0.1) | 0, 4<br>0.7 (0.8) | 0, 4.5<br>2.1 (1.1) |

both the first and second injections (27.4 (1.7) °C and 28.1 (1.6) °C, respectively). The surface body temperature of the survivor group returned to baseline within 96 hours following the second LPS injection. Surface temperature of non-survivor animals was the lowest 10.5 hours following the first injection (25.4 (1.8) °C) and 9 hours following the second injection (23.4 (1.2) °C), respectively. No significant decrease in core and surface temperatures from baseline was observed in control animals.

### 3.2.2 Body weight in survivor and non-survivor animals

In the stroke model, a decrease in body weight was observed in control animals, survivors and non-survivors (Tab. 3). In the control group, body weight was 24.8 (3.2) g at baseline decreasing to a minimum of 22.3 (2.5) g on the 2nd day post-treatment.

The lowest weight of survivor animals 21.0 (2.8) g was measured on the 2nd day following MCAo. The non-survivor group had the most profound decrease of 8.0 g from baseline (26.9 (3.7) g). The lowest mean body weight was 18.9 (2.6) g on the 4th day following MCAo. Once the minimum was reached, all groups subsequently recovered body weight until the end of the observation period.

In the sepsis model, no weight changes other than random fluctuations were observed in control animals. Body weight of the survivor group reached its minimum 54 hours following the first LPS injection (17.7 (1.5) g) and returned to baseline at 192 hours (21.7 (1.6) g, as compared to baseline weight = 21.5 (1.8) g). The lowest weight of non-survivor animals (baseline weight: 21.4 (2.0) g) was measured 96 hours after the first injection (14.4 (2.4) g).

### 3.2.3 Sickness severity score in survivor and non-survivor animals

In the stroke model, modified DeSimoni neuroscore of control animals was 0.9 (1.0) at baseline, peaked at 5.7 (4.3) on the 1st day after MCAo and subsequently decreased on the 2nd day after MCAo (Tab. 3). Non-survivors had a higher score on the 1st and 2nd day following MCAo than survivors (16.0 (9.6) and 14.8 (9.4) vs. 9.2 (5.2) and 9.1 (5.7), respectively).

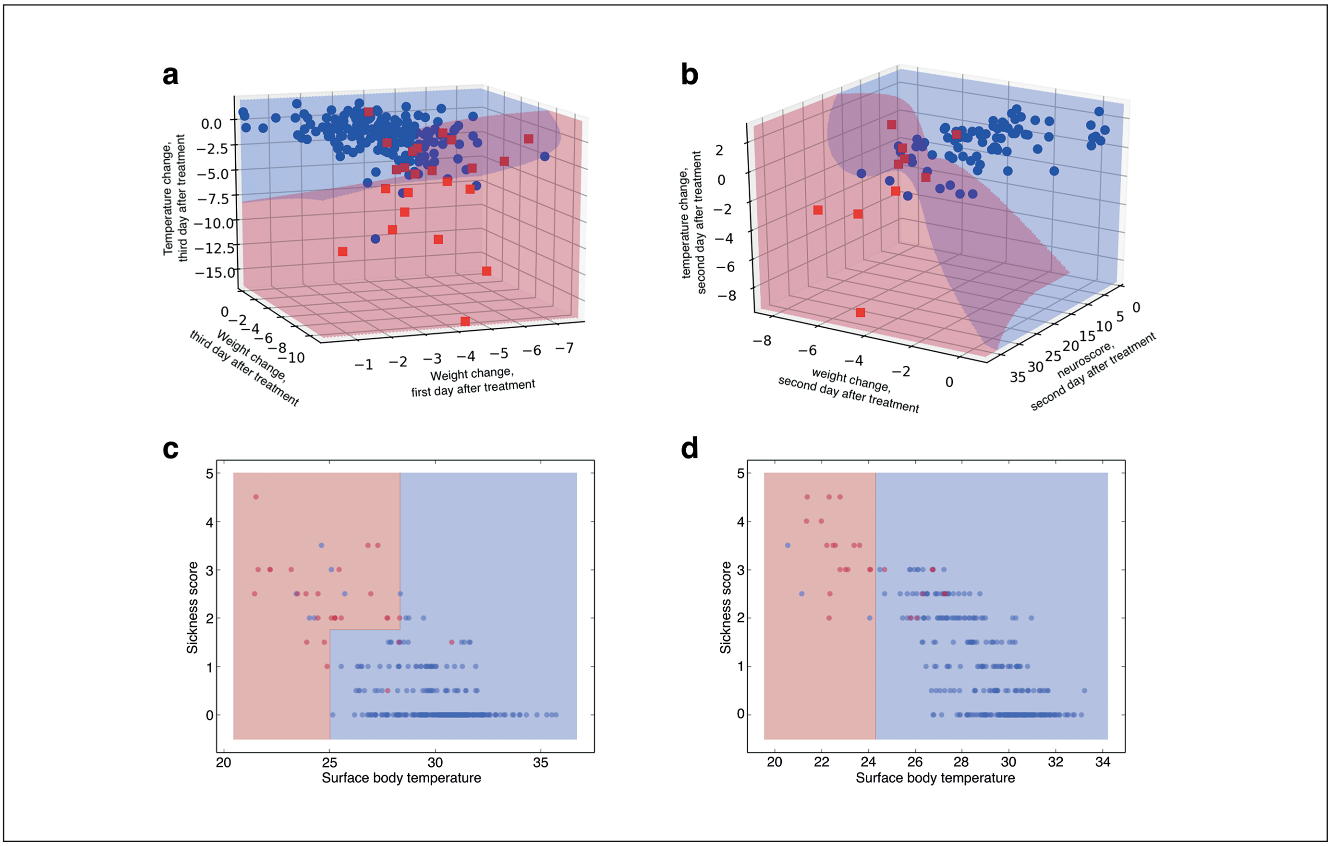

**Fig. 2: Decision boundaries determined by the machine learning model**

The earliest time points (2 and 3 days or 24 h post-treatment in the stroke or sepsis model, respectively) at which impending death could be predicted with acceptable accuracy were included. Data from 36 h post-injection in the sepsis model was plotted for comparison purposes. Parameter-model combinations leading to highest prediction accuracy were plotted. (a) Decision boundary obtained with body weight change on the 1st and 3rd day after treatment and core body temperature change on the 3rd day after treatment. (b) Decision boundary obtained with the modified DeSimoni neuroscore, body weight change and core body temperature change on the 2nd day after treatment. (c) Decision boundary obtained with surface temperature and sickness score 24 h after the first injection of LPS/saline. (d) Decision boundary obtained with surface temperature and sickness score 36 h after the first injection. Blue dot, survivor animal (control + survivor); red dot, animal euthanized upon reaching the pre-defined sickness score-based humane endpoint or died spontaneously (non-survivor); blue area, predicted survival; red area, predicted death. Gaussian Naïve Bayes (as in a and b), decision trees of depth 2 (as in c) and 1 (as in d) were used to determine the decision boundaries.

In the sepsis model, sickness scores of the survivor group peaked at 10.5 hours (1.2 (0.8)) and 12 hours (1.4 (0.9)) after the injection on day 1 and 2, respectively, then returned to baseline level within 96 hours following the first LPS injection. Sickness scores of non-survivors increased after injection day 1 and peaked at 10.5 hours after the second injection (3.3 (0.6)). The sickness scores among control animals remained unchanged (Tab. 3).

### 3.2.4 Prediction of death from physiological measures

To assess the performance of physiological and behavioral parameters such as body weight, temperature, and sickness severity score/modified DeSimoni neuroscore in death prediction, we developed an automated parameter search method to test the performance of machine learning models trained with different combinations of parameters. Apart from body weight, temperature, and sickness score/neuroscore, the absolute change per timepoint for these parameters was calculated by subtracting the baseline value from measured values at each timepoint, resulting in an additional parameter set. The two sets of parameters were used in model training.

Machine learning models were trained with individual parameters or combinations of physiological and behavioral parameters. Model performance was analyzed for time points prior to the average time of death (i.e., 3.9 (2.4) days in the stroke model and 60.5 (35.1) hours in the sepsis model, respectively). Animals in the stroke model displayed a significant gender-dependent difference in baseline body weight (male: 26.8 (2.8) g; female: 23.1 (2.9) g; $p < 0.001$, Mann-Whitney U test), thus death prediction was conducted separately for each gender.

Death as an outcome event could be predicted with considerable accuracy of 93.2% (male) or 93.0% (female) in the model of stroke and 96.2% in the model of sepsis (Tab. 4a,b), with weight change on the 1st and 3rd day after treatment, core temperature change on the 3rd day after treatment (male, stroke model), or with neuroscore, weight change, and core temperature change on the 2nd day after treatment (female, stroke model), and with surface temperature and sickness score at 24 hours after the first injection (sepsis model). Gaussian Naïve Bayes (male and female, stroke model), decision tree of a depth of 2 (with data from 24 hours after the first injection, sepsis model) and 1 (with data from 36 hours after the first injection, sepsis model) were used to identify decision boundaries shown in Figure 2.

In male mice of the stroke model, 13 out of 23 animals that died or reached predefined humane endpoint criteria at a later time point could have been euthanized earlier (t = 3 days after MCAo for euthanasia; average time of death of the 13 animals = 4.08 (1.07) days post-MCAo) while 3.3% (6 out of 181) survivors were falsely predicted to die. In female mice of the stroke model, 4 out of 10 animals that died or reached the predefined humane endpoint during the experiment could have been euthanized earlier (t = 2 days after MCAo for euthanasia; average time of death of the 4 animals = 4.25 (2.28) days post-MCAo). 3.9% (3 out of 77) animals that survived until the end of the experiment were falsely predicted to die (Fig. 2a,b).

In the sepsis model, 25 out of 28 animals could have been euthanized at an earlier time point (t = 24 hours post treatment for earlier euthanasia; average time of death of the 25 animals = 58.7 (35.0) hours post treatment) while 2.3% (6 out of 254) of LPS-treated animals that survived until the end of the experiment were falsely predicted to die (Fig. 2c,d).

*Prediction of death as an outcome event at different post-treatment time points*

By applying machine learning models trained with physiological parameters, death could be predicted within 2 or 3 days (stroke model) or 24 hours (sepsis model) after MCAo or LPS injection. In the stroke model, death could not be predicted at an acceptable level of accuracy until the 2nd (female mice) or 3rd (male mice) day post-MCAo (Tab. 4a). In the sepsis model, physiological measures obtained 12 hours after the first injection could not predict death as an outcome event due to the low general performance of the model at this time point (for details see Tab. 4b).

*Prediction of death by using single or multiple physiological measurements*

In the stroke model, adding additional physiological parameters in model training increased death prediction performance (Tab. 4a). In male mice, adding weight change on the 1st day after treatment and core temperature change on the 3rd day after treatment increased sensitivity from 0.34 (0.15) to 0.61 (0.088), precision from 0.64 (0.27) to 0.74 (0.070), and accuracy from 0.91 (0.03) to 0.93 (0.02). In female mice, using neuroscore and core temperature change on the 2nd day after treatment in addition to weight change improved sensitivity from 0.19 (0.14) to 0.69 (0.28), precision from 0.25 (0.20) to 0.83 (0.24) and accuracy from 0.86 (0.045) to 0.93 (0.030) of the trained model.

In the sepsis model, using both sickness score and surface temperature in model training improved accuracy and general performance of the prediction model (Tab. 4b). At 24 hours after treatment, using sickness score in addition to surface temperature in model training led to an increase in sensitivity from 0.65 (0.11) to 0.86 (0.12) and accuracy from 0.95 (0.01) to 0.96 (0.01), while model precision decreased slightly from 0.77 (0.17) to 0.75 (0.11).

*Interaction between the use of multiple physiological measurements and time points*

In the stroke model, due to the low performance of models trained with combinations of physiological parameters obtained on the same day, data from different post-treatment days were used in combination in model training. This approach precluded the assessment of an interaction between multiple parameters and individual time points.

In the sepsis model, using multiple physiological parameters to train the predictive model enhanced the accuracy of death prediction at individual time points. When the model was trained with data from 24 hours after the first LPS injection, using sickness score as an additional measure increased sensitivity and accuracy by 21.5% and 1.1%, respectively (Tab. 4b). No improvement in model performance was observed using data from 36 hours after the first LPS injection when sickness scores were included in model training (Tab. 4b).

## 4 Discussion

Our study revealed three main findings: Firstly, the systematic review demonstrated remarkable heterogeneity of humane endpoints even within the same animal model due to lack of systematic assessment, protocol standardization, and/or ambiguous or incomplete description of results. This illustrates a distinct challenge in adopting humane endpoints from the literature and highlights the need for researchers to tailor humane endpoints based on the currently available evidence. Secondly, using data from mouse models of stroke and sepsis, we found that machine learning by means of an automated search for predictive parameters, parameter combinations, and models and hyperparameters, can be used to accurately determine endpoint criteria and cut-off threshold values across models. Thirdly, when we applied these criteria retrospectively to the available derivation cohort datasets, we found that a large number of animals could have been euthanized at earlier time points in both stroke and sepsis models, thus potentially reducing otherwise unnecessary suffering.

### 4.1 Systematic review

In this study, we reviewed 34 mouse studies using humane endpoints based on body temperature, body weight and/or sickness severity score (Tab. 5). We found that temperature-based endpoints are commonly applied both in acute and chronic disease

**Tab. 4: Death prediction with single or multiple parameters at different time points**
(A) Middle cerebral artery occlusion (MCAo) stroke model. Prediction model, Gaussian Naïve Bayes. (B) Lipopolysaccharide (LPS)-induced sepsis model. Prediction model, decision tree with a depth = 2 (24 h) and a depth = 1 (36 h). Decision tree of depth 2 was not used for 36 h after the first injection due to overfitting. 3-fold stratified cross-validation was used to evaluate the performance of the trained model. Only performance of the most predictive parameters and model combinations is shown. Data shown are means (SD) of scores obtained from the 2- (stroke model) or 3- (sepsis model) fold cross-validation.

**A**

| **Gender** | male (n = 204, n(dead) = 23) | | | female (n = 87, n(dead) = 10) | | |
|---|---|---|---|---|---|---|
| **Parameters** | weight change on the 3rd day after treatment | weight change on the 1st day after treatment;<br>weight change on the 2nd day after treatment;<br>core temperature change on the 2nd day after treatment | weight change on the 1st day after treatment;<br>weight change on the 3rd day after treatment;<br>core temperature change on the 3rd day after treatment | weight change on the 2nd day after treatment | neuroscore on the 1st day after treatment;<br>weight change on the 1st day after treatment;<br>core temperature change on the 1st day after treatment | neuroscore on the 2nd day after treatment;<br>weight change on the 2nd day after treatment;<br>core temperature change on the 2nd day after treatment |
| **Sensitivity (recall)** | 0.339 (0.148) | 0.482 (0.081) | 0.613 (0.088) | 0.194 (0.142) | 0.361 (0.307) | 0.694 (0.275) |
| **Precision** | 0.644 (0.274) | 0.667 (0.272) | 0.738 (0.070) | 0.25 (0.204) | 0.417 (0.312) | 0.833 (0.236) |
| **Accuracy** | 0.907 (0.026) | 0.902 (0.046) | 0.932 (0.018) | 0.863 (0.045) | 0.896 (0.029) | 0.930 (0.030) |
| **Averaged score** | 0.630 | 0.684 | 0.761 | 0.436 | 0.558 | 0.819 |

**B**

| **Time** | t = 12 h post-treatment (n = 152, n(dead) = 14) | | t = 24 h post-treatment (n = 345, n(dead) = 28) | | t = 36 h post-treatment (n = 342, n(dead) = 25) | |
|---|---|---|---|---|---|---|
| **Parameters** | surface temperature | surface temperature, sickness score | surface temperature | surface temperature, sickness score | surface temperature | surface temperature, sickness score |
| **Sensitivity (recall)** | 0 | 0 | 0.648 (0.114) | 0.863 (0.124) | 0.685 (0.092) | 0.685 (0.092) |
| **Precision** | 0 | 0 | 0.768 (0.165) | 0.747 (0.106) | 0.806 (0.142) | 0.806 (0.142) |
| **Accuracy** | 0.908 (0.009) | 0.908 (0.009) | 0.951 (0.004) | 0.962 (0.011) | 0.962 (0.004) | 0.962 (0.004) |
| **Averaged score** | 0.303 | 0.303 | 0.789 | 0.857 | 0.818 | 0.818 |

models due to their objectivity and ease of measurement. However, we found considerable variations in temperature cut-off values between studies even within the same animal model (Molins et al., 2012; Beyer et al., 2009; Adamson et al., 2013; Bast et al., 2004; Cates et al., 2014; Hankenson et al., 2013; Trammell and Toth, 2011; Nemzek et al., 2004; Dellavalle et al., 2014; Hunter et al., 2014; Warn et al., 2003; Kort et al., 1998; Vlach et al., 2000; Ray et al., 2010), which can at least partly be explained by variations in ambient temperature.

Ambient temperature is an important factor contributing to differences between an animal's core and surface temperature. The lower the ambient temperature, the lower an animal's surface temperature, while the animal's core temperature stays constant as long as thermoregulatory responses are intact (Kurz, 2008). In addition, measurement location and handling stress may contribute to differences in cut-off values between studies. Some authors used restraining devices for probe-based surface temperature acquisition. However, stress results in activation of the sympathetic nervous system, which in turn leads to increased thermogenesis and vasoconstriction of skin vessels, resulting in an increase in body temperature within seconds of being restrained (Vianna and Carrive, 2005). Therefore, body temperature measurements could be confounded by repeated handling (Cabanac and Briese, 1992).

**Tab. 5: Summary of endpoint criteria, advantages and disadvantages of weight-, body temperature-, and severity score-based humane endpoints**

| Humane endpoint | Studies included | Endpoint | Advantages | Disadvantages |
|---|---|---|---|---|
| **Weight-based** | 13 | 15-25% of baseline body weight | – easy administration<br>– high objectivity | – poor performance in acute disease models<br>– high handling stress |
| **Core temperature-based** | 11 | 23.4-36°C, 31.55 (4.7) °C | – high objectivity<br>– high accuracy<br>– continuous monitoring<br>– low handling stress | – high variance due to an animal's thermo-regulatory responses and handling stress |
| **Surface temperature-based** | 4 | 28.8-30°C, 29.7 (0.6) °C | – high objectivity<br>– high accuracy<br>– low handling stress | – high variance due to an animal's thermo-regulatory responses and handling stress |
| **Severity score-based** | 14 | Multiple criteria | – easy administration<br>– simplified classification of physiological states<br>– systematic documentation | – requires familiarity<br>– inter-observer variability<br>– time-consuming<br>– high handling stress |

Humane endpoints based on rapid (over a few days) or gradual (over extended periods of time leading to emaciation) weight loss relative to baseline are easy to adopt and are widely applied. However, weight-based endpoints are suboptimal in highly acute models of disease (i.e., circulatory shock) due to an animal's rapid deterioration which may precede weight loss (Louie et al., 1997; Krarup et al., 1999; Nemzek et al., 2004). In addition, true weight loss may be masked by debilitating conditions such as ascites or tumor growth (Nemzek et al., 2004).

Sickness severity scores serve as a simplified classification of the physiological state of an animal, allowing systematic documentation of disease progression and humane endpoint evaluation. However, manual scoring suffers from subjectivity and is prone to high degrees of inter-observer bias (Morton, 2000).

Another factor contributing to high study heterogeneity is the lack of standardized schedules for animal inspection. Even for identical animal models, authors rarely used measurement schedules which were consistent with previously published data (Nunamaker et al., 2013a,b). Furthermore, only a minority of studies (9 out of 34) described the exact times relative to baseline and/or experimental intervention when temperature, body weight, and/or sickness score values were taken. Other variables potentially adding to study heterogeneity but only described by few studies include environmental factors such as the presence and type of bedding (described in 17 out of 34 studies) and number of cage mates (described in 22 out of 34 studies; Gordon, 2004; Gordon et al., 1998), as well as animal-specific factors such as strain, genotype, sex and developmental stage (described in all reviewed studies; Sanchez-Alavez et al., 2011; Trammell and Toth, 2011).

To counter the uncertainty in endpoint evaluation caused by variation of individual parameters, 15 of the reviewed studies used a combination of more than one humane endpoint criterion. Body weight was most commonly applied in combination with additional criteria (13 out of 13 studies used additional criteria) while sickness scores were mostly used independently (3 out of 13 studies used additional criteria). In 2 studies that evaluated the use of a composite score derived from other parameters (i.e., the product of weight x body temperature), a higher prediction accuracy was observed than when assessing parameters individually (Trammell and Toth, 2011; Ray et al., 2010). Authors used various metrics to assess the performance of humane endpoints, which often precludes direct between-study comparison. The three most common metrics were sensitivity (n = 7), specificity (n = 5), and percentage/number of mice exhibiting certain criteria/signs (n = 5). In 12 out of the 34 studies, humane endpoints were applied without being evaluated for reliability and/or performance or without endpoint evaluation being reported by the authors. Therefore, researchers may fail to appreciate the validity and/or reproducibility of humane endpoint criteria.

Taken together, traditional approaches in monitoring disease progression and predicting death suffer from high degrees of study heterogeneity, which confounds identification of cut-off values that can be applied to more than one study.

### 4.2 Machine learning-based death prediction

In an exploratory approach, we used machine learning as an alternative method for determining humane endpoints, which enabled us to identify case-specific cut-off values even across animal models without a fundamental change in methodology. Using body weight, sickness severity scores, and surface or core temperature data (for the sepsis or stroke model, respectively) from previously published studies and unpublished results, we trained a machine learning model for case-specific death prediction. First, we identified the parameter combinations that led to a high accuracy in detecting animals at higher risk of death. We then determined the cut-off values and assessed their performance using standardized metrics. We found that 17 out of 33 (stroke model) and 25 out of 28 (sepsis model) animals that were euthanized or died at later timepoints during the experiments could have been euthanized 1.08 days (stroke model, male), 2.25 days (stroke model, female) or 1.45 days (sepsis model) earlier if endpoints determined by machine learning had been applied.

To our knowledge, this is the first study using a machine learning approach to systematically determine humane endpoints in mouse models of acute disease and there is no previous data to compare our results to.

Potential advantages of a machine learning-based approach for humane endpoint evaluation include improved standardization and comparability of results as identical metrics can be applied across studies. In addition, machine learning in this setting requires little expert knowledge and is relatively easy to apply. Once a model is trained with sufficient data, animal technicians and investigators can run the model with a simple command line tool to determine whether an animal has an increased risk of death, for example by visualizing decision boundaries (e.g., as shown in Fig. 2). Thus, machine learning may constitute a promising tool to improve the refinement of humane endpoints in animal studies of acute disease. However, the availability of data which can be used to train and/or validate machine learning models to evaluate and refine humane endpoints across disease models is limited as only very few authors choose to make their raw data available in open data repositories.

### 4.3 Limitations of the present study

In the systematic review, we excluded research articles in languages other than English and results published in the form of conference abstracts, posters and talks. This introduces a potential source of publication and result bias. A particular concern for the machine learning-based analysis was missing data and a small number of animals which reached a humane endpoint criterion. Nevertheless, for humane endpoint evaluation small datasets and missing data represent the real-world scenario. One solution is to increase the frequency of inspections during critical phases, not only to ensure that disease progression is properly monitored, but to obtain a larger dataset for training death prediction models. To this end, some authors suggest the use of biotelemetry systems with implanted transmitters, which allow continuous data acquisition (Leon et al., 2005). Lastly, we applied machine learning-derived endpoint criteria to the derivation cohort, which may introduce results bias. Although a stratified cross-validation was used to optimize generalizability, it would be beneficial for future studies to include an independent retrospective or prospective validation cohort to assess predictive model performance.

### 4.4 Conclusion

The degree of heterogeneity between studies using body temperature, weight and sickness scores to determine or evaluate humane endpoint criteria was high. This may preclude authors from adopting appropriate cut-off values from previously published studies and underscores the necessity for researchers to tailor the humane endpoints to the animal model and experimental design being used. In an approach aimed at enhancing the evaluation and application of humane endpoints using machine learning, we identified humane endpoints that would have allowed earlier euthanasia of animals, thus potentially reducing an animal's distress, suffering, and pain. Although the method still requires further validation, this exploratory study showed that machine learning-based humane endpoint criteria have the potential to be applied across various disease models. This may contribute to a more comprehensive approach in determining humane endpoints promoting the systematic application of 3R principles.

## References

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.

Aldred, A. J., Cha, M. C. and Meckling-Gill, K. A. (2002). Determination of a humane endpoint in the L1210 model of murine leukemia. *Contemp Top Lab Anim Sci 41*, 24-27.

Arranz-Solis, D., Aguado-Martinez, A., Muller, J. et al. (2015). Dose-dependent effects of experimental infection with the virulent Neospora caninum Nc-Spain7 isolate in a pregnant mouse model. *Vet Parasitol 211*, 133-140. doi:10.1016/j.vetpar.2015.05.021

Bast, D. J., Yue, M., Chen, X. et al. (2004). Novel murine model of pneumococcal pneumonia: Use of temperature as a measure of disease severity to compare the efficacies of moxifloxacin and levofloxacin. *Antimicrob Agents Chemother 48*, 3343-3348. doi:10.1128/aac.48.9.3343-3348.2004

Benedito, R., Roca, C., Sörensen, I. et al. (2009). The notch ligands Dll4 and Jagged1 have opposing effects on angiogenesis. *Cell 137*, 1124-1135. doi:10.1016/j.cell.2009.03.025

Beyer, N. H., Kogutowska, E., Hansen, J. J. et al. (2009). A mouse model for ricin poisoning and for evaluating protective effects of antiricin antibodies. *Clin Toxicol (Phila) 47*, 219-225. doi:10.1080/15563650802716521

Cabanac, A. and Briese, E. (1992). Handling elevates the colonic temperature of mice. *Physiol Behav 51*, 95-98. doi:10.1016/0031-9384(92)90208-j

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

Chappell, M. G., Koeller, C. A. and Hall, S. I. (2011). Differences in postsurgical recovery of CF1 mice after intraperitoneal implantation of radiotelemetry devices through a midline or flank surgical approach. *J Am Assoc Lab Anim Sci 50*, 227-237.

Dellavalle, B., Kirchhoff, J., Maretty, L. et al. (2014). Implementation of minimally invasive and objective humane endpoints in the study of murine Plasmodium infections. *Parasitology 141*, 1621-1627. doi:10.1017/s0031182014000821

Dirnagl, U. and Members of the MCAO-SOP Group. (2012). Standard operating procedures (SOP) in experimental stroke research: SOP for middle cerebral artery occlusion in the mouse. *Nat Prec.* doi:10.1038/npre.2012.3492.3

Donath, S., An, J., Lee, S. L. L. et al. (2016). Interaction of ARC and Daxx: A novel endogenous target to preserve motor function and cell loss after focal brain ischemia in mice. *J Neurosci 36*, 8132-8148. doi:10.1523/jneurosci.4428-15.2016

Emmrich, J. V., Neher, J. J., Boehm-Sturm, P. et al. (2017). Stage 1 registered report: Effect of deficient phagocytosis on neuronal survival and neurological outcome after temporary middle cerebral artery occlusion (tMCAo). *F1000Res 6*, 1827. doi:10.12688/f1000research.12537.1

Faller, K. M. E., McAndrew, D. J., Schneider, J. E. and Lygate, C. A. (2015). Refinement of analgesia following thoracotomy and experimental myocardial infarction using the mouse grimace scale. *Exp Physiol 100*, 164-172. doi:10.1113/expphysiol.2014.083139

Franco, N. H., Correia-Neves, M. and Olsson, I. A. S. (2012). How "humane" is your endpoint? – Refining the science-driven approach for termination of animal studies of chronic infection. *PLoS Pathog 8*, e1002399. doi:10.1371/journal.ppat.1002399

Glerup, S., Olsen, D., Vaegter, C. B. et al. (2014). SorCS2 regulates dopaminergic wiring and is processed into an apoptotic two-chain receptor in peripheral glia. *Neuron 82*, 1074-1087. doi:10.1016/j.neuron.2014.04.022

Gordon, C. J., Becker, P. and Ali, J. S. (1998). Behavioral thermoregulatory responses of single-and group-housed mice. *Physiol Behav 65*, 255-262. doi:10.1016/S0031-9384(98)00148-6

Gordon, C. J. (2004). Effect of cage bedding on temperature regulation and metabolism of group-housed female mice. *Comp Med 54*, 63-68.

Han, S., Taralova, E., Dupre, C. and Yuste, R. (2018). Comprehensive machine learning analysis of Hydra behavior reveals a stable basal behavioral repertoire. *eLIFE 7*, e32605. doi:10.7554/elife.32605

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.

Hoffmann, C. J., Harms, U., Rex, A. et al. (2015). Vascular signal transducer and activator of transcription-3 promotes angiogenesis and neuroplasticity long-term after stroke. *Circulation 131*, 1772-1782. doi:10.1161/circulationaha.114.013003

Huet, O., Ramsey, D., Miljavec, S. et al. (2013). Ensuring animal welfare while meeting scientific aims using a murine pneumonia model of septic shock. *Shock 39*, 488-494. doi:10.1097/shk.0b013e3182939831

Hunter, J. E., Butterworth, J., Perkins, N. D. et al. (2014). Using body temperature, food and water consumption as biomarkers of disease progression in mice with Eμ-myc lymphoma. *Br J Cancer 110*, 928-934. doi:10.1038/bjc.2013.818

Husmann, K., Arlt, M. J. E., Jirkof, P. et al. (2015). Primary tumour growth in an orthotopic osteosarcoma mouse model is not influenced by analgesic treatment with buprenorphine and meloxicam. *Lab Anim 49*, 284-293. doi:10.1177/0023677215570989

Kabra, M., Robie, A. A., Rivera-Alba, M. et al. (2013). JAABA: Interactive machine learning for automatic annotation of animal behavior. *Nat Methods 10*, 64-67. doi:10.1038/nmeth.2281

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

Koch, S., Mueller, S., Foddis, M. et al. (2017). Atlas registration for edema-corrected MRI lesion volume in mouse stroke models. *J Cereb Blood Flow Metab 39*, 313-323. doi:10.1177/0271678x17726635

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

Krarup, A., Chattopadhyay, P., Bhattacharjee, A. K. et al. (1999). Evaluation of surrogate markers of impending death in the galactosamine-sensitized murine model of bacterial endotoxemia. *Comp Med 49*, 545-550.

Kurz, A. (2008). Physiology of thermoregulation. *Best Pract Res Clin Anaesthesiol 22*, 627-644. doi:10.1016/j.bpa.2008.06.004

Leon, L. R., DuBose, D. A. and Mason, C. W. (2005). Heat stress induces a biphasic thermoregulatory response in mice. *Am J Physiol Regul Integr Comp Physiol 288*, R197-R204. doi:10.1152/ajpregu.00046.2004

Louie, A., Liu, W., Liu, Q.-F. et al. (1997). Predictive value of several signs of infection as surrogate markers for mortality in a neutropenic guinea pig model of Pseudomonas aeruginosa sepsis. *Lab Anim Sci 47*, 617-623.

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

Miller, A., Burson, H., Söling, A. and Roughan, J. (2016). Welfare assessment following heterotopic or orthotopic inoculation of bladder cancer in C57BL/6 mice. *PLoS One 11*, e0158390. doi:10.1371/journal.pone.0158390

Miller, D. S., Kok, T. and Li, P. (2013). The virus inoculum volume influences outcome of influenza A infection in mice. *Lab Anim 47*, 74-77. doi:10.1258/la.2012.011157

Molins, C. R., Delorey, M. J., Young, J. W. et al. (2012). Use of temperature for standardizing the progression of Francisella tularensis in mice. *PLoS One 7*, e45310.

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

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., Anderson, R. J., Artwohl, J. E. et al. (2013a). Predictive observation-based endpoint criteria for mice receiving total body irradiation. *Comp Med 63*, 313-322.

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

OECD (2000). Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for

Experimental Animals Used in Safety Evaluation. *Series on Testing and Assessment No. 19*. OECD Publishing, Paris. doi:10.1787/9789264078376-en

Passman, A. M., Strauss, R. P., McSpadden, S. B. et al. (2015). A modified choline-deficient, ethionine-supplemented diet reduces morbidity and retains a liver progenitor cell response in mice. *Dis Model Mech 8*, 1635-1641. doi:10.1242/dmm.022020

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.

Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011). Scikit-learn: Machine learning in python. *J Mach Learn Res 12*, 2825-2830. doi:10.1002/9781119557500.ch5

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.

Sanchez-Alavez, M., Alboni, S. and Conti, B. (2011). Sex-and age-specific differences in core body temperature of C57BL/6 mice. *Age 33*, 89-99. doi:10.1007/s11357-010-9164-6

Sand, C. A., Starr, A., Nandi, M. and Grant, A. D. (2015). Blockade or deletion of transient receptor potential vanilloid 4 (TRPV4) is not protective in a murine model of sepsis. *F1000Res 4*, 93. doi:10.12688/f1000research.6298.1

Skarnes, W. C., Rosen, B., West, A. P. et al. (2011). A conditional knockout resource for the genome-wide study of mouse gene function. *Nature 474*, 337-342. doi:10.1038/nature10163

Slezak, M., Göritz, C., Niemiec, A. et al. (2007). Transgenic mice for conditional gene manipulation in astroglial cells. *Glia 55*, 1565-1576. doi:10.1002/glia.20570

Solomon, J. A., Tarnopolsky, M. A. and Hamadeh, M. J. (2011). One universal common endpoint in mouse models of amyotrophic lateral sclerosis. *PLoS One 6*, e20582. doi:10.1371/journal.pone.0020582

Stoica, L., Todeasa, S. H., Cabrera, G. T. et al. (2016). Adeno-associated virus-delivered artificial microRNA extends survival and delays paralysis in an amyotrophic lateral sclerosis mouse model. *Ann Neurol 79*, 687-700. doi:10.1002/ana.24618

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

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.

Vianna, D. M. and Carrive, P. (2005). Changes in cutaneous and body temperature during and after conditioned fear to context in the rat. *Eur J Neurosci 21*, 2505-2512. doi:10.1111/j.1460-9568.2005.04073.x

Vlach, K. D., Boles, J. W. and Stiles, B. G. (2000). Telemetric evaluation of body temperature and physical activity as predictors of mortality in a murine model of staphylococcal enterotoxic shock. *Comp Med 50*, 160-166.

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

Weismann, C. M., Ferreira, J., Keeler, A. M. et al. (2015). Systemic AAV9 gene transfer in adult GM1 gangliosidosis mice reduces lysosomal storage in CNS and extends lifespan. *Hum Mol Genet 24*, 4353-4364. doi:10.1093/hmg/ddv168

Wright, A. J. and Phillpotts, R. J. (1998). Humane endpoints are an objective measure of morbidity in Venezuelan encephalomyelitis virus infection of mice. *Arch Virol 143*, 1155-1162. doi:10.1007/s007050050363

Yamaguchi, S., Naoki, H., Ikeda, M. et al. (2018). Identification of animal behavioral strategies by inverse reinforcement learning. *PLoS Comput Biol 14*, e1006122. doi:10.1371/journal.pcbi.1006122

## Conflict of interest

The authors declare that they have no conflicts of interest.

## Acknowledgements

This work was supported by Berlin-Brandenburg School for Regenerative Therapies, Theracur Foundation, German Research Foundation (grant numbers: HO 5177/3-1, HA 5741/5-1, EM 252/2-1), German Federal Ministry of Education and Research, Berlin Institute of Health and Einstein Stiftung Berlin (grant number: 2014-223). Dr Hoffmann is a participant in the BIH Charité Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health.