Aus der Klinik für Neurologie (Abteilung Experimentelle Neurologie)
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

# Sepsis-associated cognitive dysfunction: An investigation using stress-free, automated behavioral tests

# Sepsis-assoziierte kognitive Dysfunktion: Eine Untersuchung mit stressfreien, automatisierten Verhaltenstests

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Jie Mei

aus Jilin, China

Datum der Promotion: 21.06.2020

## Table of Contents

## List of Figures

## List of Tables

## List of Abbreviations

| | |
|---|---|
| Cd11b | Cluster of differentiation molecule 11b |
| CNS | Central nervous system |
| CR3 | Complement receptor 3 |
| cRGD | Cyclic arginine-glycine-aspartate |
| Gas6 | Growth arrest specific gene 6 |
| ICU | Intensive care units |
| LPS | Lipopolysaccharide |
| MCAo | Middle cerebral artery occlusion |
| Mertk | Mer receptor tyrpsine kinase |
| Mfge8 | Milk fat globule EGF-like factor 8 |
| PBS | Physiological phosphate-buffered saline |
| PS | Phosphatidylserine |
| RAM | Radial arm maze |
| RFID | Radio-frequency identification |

## Abstract

Survivors of medical and surgical intensive care units (ICUs) are at high risk for long-lasting cognitive impairments. Among critical illnesses that can induce cognitive deficits, sepsis is commonly regarded as the most frequent and severe cause, considering one out of three survivors of sepsis is discharged from hospital with severe *de novo* cognitive impairment. Chronic neuroinflammation, diffuse cerebral damage and neuronal death are considered primary correlates in the development of long-term cognitive deficits. Recent studies have suggested sepsis can cause inflammatory activation of the microglia, which lead to microglial phagocytosis of stressed but viable neurons.

To establish a mouse model of sepsis, lipopolysaccharide (LPS) at a dose of 1.5 mg/kg was injected in C57BL/6 mice and homozygous knockout and wildtype mice that are deficient for Mertk, Cd11b and Mfge8. Immediate sickness behavior and long-term cognitive functions of animals were analyzed to assess the effects of phagocytic deficiency and peptide treatments on cognitive deficits. To best meet the requirements of animal welfare by minimizing repetitive handling, water and/or food restriction and unnecessary suffering, we have also investigated the applicability of a fully automatized 8-arm radial arm maze (RAM) and machine learning-based humane endpoint determination.

LPS-injected animals have displayed (a) characteristics of sickness behavior immediately following the injections, and (b) cognitive deficits after the one-month recovery period. Animals deficient for Mertk, Cd11b or Mfge8 have displayed greater learning performances during place learning, place reversal and avoidance conditioning. Treatment effects of Cilengitide or cRGD were observed in sucrose preference, avoidance conditioning and the early stage of place learning. In the meantime, using humane endpoints determined with machine learning models, mice of both stroke and sepsis model that are at higher risk of death could be detected at a high accuracy. Mice up to 18 months of age have shown efficient spatial learning in both working memory and combined working/reference memory paradigms in the automated 8-arm RAM without food and/or water restriction.

With a mouse model of sepsis, alleviation of long-term cognitive deficits could be observed in phagocytic deficient animals and Cilengitide- or cRGD-treated animals, which might offer an explanation of underlying mechanisms of long-term cognitive deficits following systemic inflammation. Minimized suffering for animals and improved reproducibility of experimental outcomes were possible using machine learning-based endpoint determination and automated behavioral testing systems, respectively.

## Zusammenfassung

Überlebende von chirurgischen und konservativen Intensivstationen haben ein hohes Risiko für lang anhaltende kognitive Defizite. Die Sepsis gilt als häufigste und schwerwiegendste der kritischen Erkrankungen, die zu kognitiven Defiziten führen können. Einer von drei Überlebenden einer Sepsis wird mit schwerer, neu aufgetretener kognitiver Dysfunktion aus dem Krankenhaus entlassen. Chronische Neuroinflammation, diffusive zerebrale Schädigungen und neuronaler Zelltod werden als die primären Korrelate in der Entwicklung lang anhaltender kognitiver Defizite angesehen. Neuere Studien deuten darauf hin, dass Sepsis eine inflammatorische Aktivierung von Mikroglia auslöst, die zu Phagozytose gestresster, aber funktionsfähiger Neurone führt.

Um ein Mausmodell der Sepsis zu etablieren, wurden Lipopolysaccharide (LPS) in C57BL/6-Mäuse und homozygote Knockout- und Wildtyp-Mäuse mit Mertk-, CD11b- und Mfge8-Defizienz in einer Dosierung von 1,5mg/kg injiziert. Das akute Krankheitsverhalten und langzeitige kognitive Funktionen der Tiere wurden analysiert, um die Effekte von phagozytotischer Defizienz und Behandlung mit Peptiden auf kognitive Defizite zu untersuchen. Um die Anforderungen an das Tierwohl durch Reduzierung von Handling, Wasser- und/oder Nahrungsentzug und unnötigem Leid optimal zu erfüllen, untersuchten wir außerdem die Anwendbarkeit eines vollständig automatisierten 8-Arm-Radial Arm Mazes und von machine learning-basierter Bestimmung von Abbruchkriterien (humane endpoints).

Tiere mit LPS-Injektion zeigten (a) Charakteristika von Krankheitsverhalten unmittelbar nach der Injektion und (b) kognitive Defizite nach der einmonatigen Erholungsperiode. Tiere mit Mertk-, CD11b- und Mfge8-Defizienz präsentierten bessere Lernleistungen bezüglich Place Learning, Place Reversal und Avoidance Conditioning. Behandlungseffekte von Cilengitid oder cRGD konnten bezüglich Sucrose Preference, Avoidance Conditioning und in der frühen Phase des Place Learning beobachtet werden. Bei der Nutzung von Machine Learning-Modellen, die Abbruchkriterien bestimmten, zeigte sich, dass Mäuse im Schlaganfall- und im Sepsismodell mit einem höheren Todesrisiko mit hoher Genauigkeit erkannt werden konnten. Mäuse bis zum Alter von 18 Monaten zeigten effizientes räumliches Lernen im Paradigma für das Arbeitsgedächtnis und im kombinierten Paradigma für das Arbeits- und Referenzgedächtnis im automatisierten 8-Arm-Radial Arm Maze ohne Nahrungs- oder Wasserentzug.

In einem Mausmodell der Sepsis beobachteten wir eine Verminderung langzeitiger kognitiver Defizite in phagozyten-defizienten und in Cilengitid- oder cRDG-behandelten Tieren, was eine Erklärung für die Mechanismen, die langzeitigen kognitiven Defiziten zugrunde liegen, bieten könnte. Minimiertes Leid für die Tiere und verbesserte Reproduzierbarkeit von experimentellen Ergebnissen waren möglich durch die Benutzung machine learning-basierter Bestimmung von Abbruchkriterien und automatisierter Verhaltenstestung.

# 1. Introduction

## 1.1. Motivation and clinical relevance

Survivors of intensive care units (ICUs) are at high risk for long-term, *de novo* cognitive deficits of memory, attention and executive functions (Pandharipande *et al.,* 2013). Failure in regaining full cognitive functions has affected 30-70% of ICU survivors at one year after discharge. Among critical illnesses that lead to cognitive deficits, sepsis is regarded as the most common cause (Widmann and Heneka, 2014). Magnetic resonance imaging studies have associated post-sepsis white matter disruption of the frontal cortex and hippocampus with long-term cognitive deficits in human survivors (Semmler *et al.,* 2013) and studies on human post-mortem brains have identified an association between systematic infection and microglial activation (Lemstra *et al.,* 2007). Therefore, it would be crucial to understand the role of microglial activation in neuronal loss and associated cognitive alterations.

## 1.2. Microglial phagocytosis and phagoptosis

Microglia are specialized macrophages which are responsible for phagocytic removal of pathogens, cell debris and dying cells, thus play an essential role in immune defense in the central nervous system (CNS). For a cell to be removed, an “eat-me” signal need to be displayed on its surface (Ravichandran, 2011; Figure 1.1). Phagocytic removal of neurons is commonly considered as a beneficial process which only occurs after neuronal death (Neher *et al.,* 2013). Nonetheless, recent findings have revealed possible removal of viable neurons by phagocytosis ("phagoptosis”; Brown and Neher, 2012). Under inflammatory conditions, phagoptosis could lead to stressed viable neurons to be executed (Fricker *et al.,* 2012).

Upon inflammatory activation using lipopolysaccharide (LPS), reversible exposure of a “eat-me” signal phosphatidylserine (PS) induced by activated microglia is observed on the surface of viable neurons, leading to possible removal (Neher *et al.,* 2013). When microglia are present, PS externalization eventually lead to neuronal death. In the absence of microglia, PS externalization could be reversed (Neher *et al.,* 2013).

## 1.3. Phagocytic deficiency, inhibition of phagocytic signaling pathways and possible alleviation of neuronal loss

Neuronal PS could be recognized through microglial transmembrane receptors (direct recognition) or binding of soluble opsonins (indirect recognition), including milk fat globule EGF-like factor 8 protein (Mfge8), growth arrest specific gene 6 (Gas6) and Protein S. Gas6 and Protein S bind to exposed PS on neurons and are recognized by the membrane protein Mer receptor tyrosine kinase (Mertk). Opsonins such as complement components C1q and C3 play a central role in synaptic pruning or pathological situations (Schafer *et al.,* 2012). C1q can also induce the conversion of C3 to C3b, and consequently recognition of the opsonized neuronal structure by complement receptor 3 (CR3), consisting of a cluster of differentiation molecule 11b (CD11b) and CD18. Microglial phagocytosis of the "eat-me" signal exposing neurons will occur as a result of activation of lipoprotein receptor-related protein or CR3 (Linnartz *et al.,* 2012).

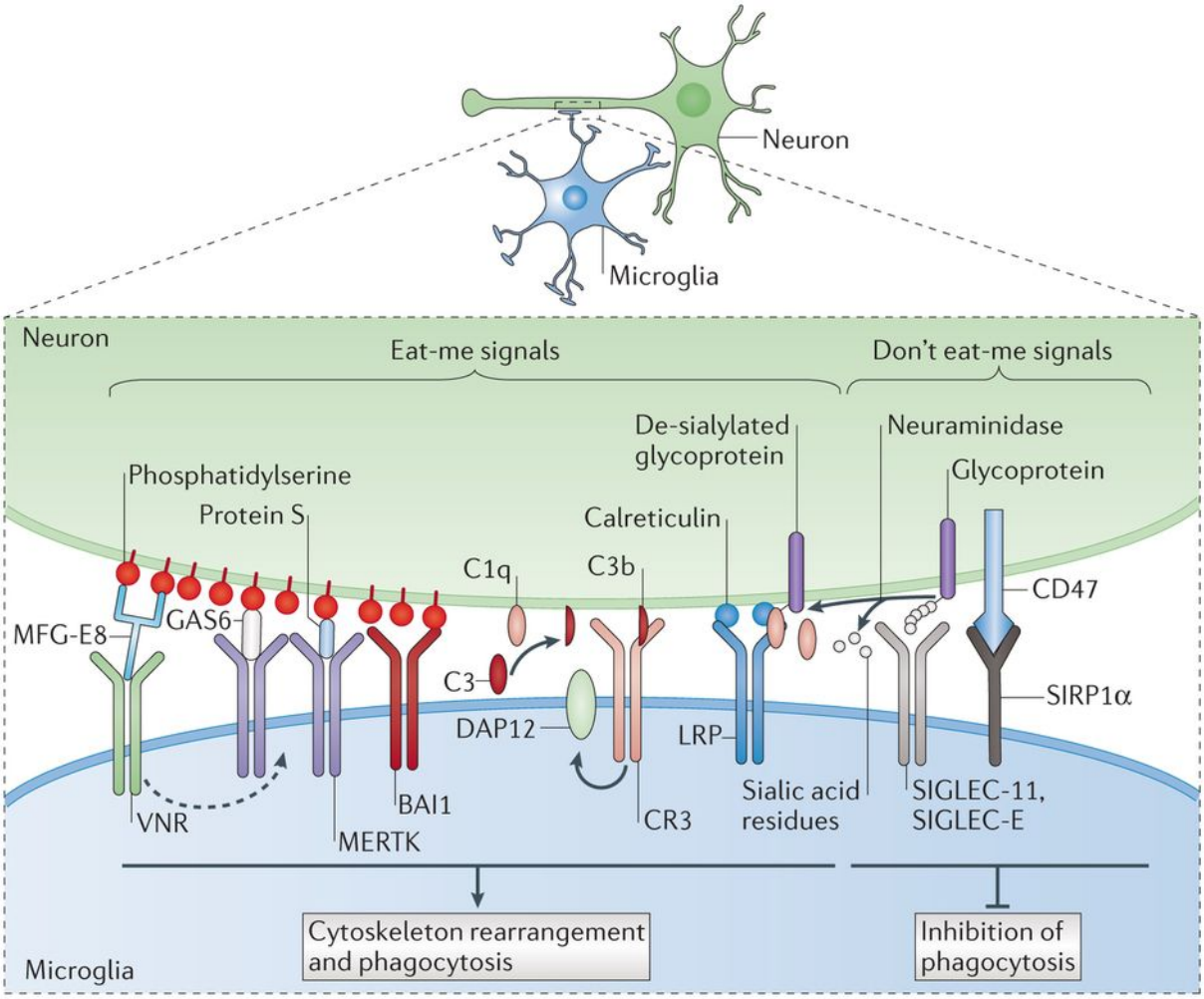


**Figure 1.1:** "Eat-me" signals, "don't eat-me" signals and signaling pathways in the regulation of phagocytosis of neurons and neuronal structures (Brown and Neher, 2014).

Previous study showed reduced neuronal loss and improved neuronal survival in LPS-injected Mfge8 knockout mice (Fricker *et al.,* 2012). In a rodent model of mild focal cerebral ischemia, mice deficient for Mfge8 or Mertk showed reduced brain atrophy and preserved sensorimotor functions along with effectively blocked phagocytosis (Neher *et al.,* 2013). To understand the effect of phagocytic deficiency on long-term cognitive abilities, in this study, animals deficient for Mertk, Cd11b and Mfge8 were used and compared with wildtype animals.

It has been shown that inhibition of phagocytic signaling pathways may lead to increased neuronal survival following focal cerebral ischemia (Emmrich *et al.,* 2017). In this study, animals were treated with Cilengitide for inhibition of phagocytic signaling, or inactive control peptide cRGD. Performance in cognitive tasks of Cilengitide-, cRGD-treated and untreated animals was assessed to examine long-term behavioral outcomes of inhibition of phagocytic signaling pathways.

### 1.4. Mouse model of LPS-induced sepsis and sickness behavior

Sickness behavior is defined as the immediate behavioral responses to pro-inflammatory mediators acting on the brain, which could be induced with LPS in a systemic inflammation model. In rodent models of acute illnesses, sickness behavior including decreased motor activity, lethargy, reduced food and water intake, reduced response to stimulus and altered cognitive functions could be observed (Dantzer *et al.,* 2008).

A primary component of sickness behavior during LPS administration is short-term behavioral alterations that only last for hours or days. Long-term affective, behavioral and cognitive alterations may occur after a recovery period (Anderson *et al.,* 2015). Previous studies have identified neuronal and synaptic loss associated with long-term cognitive deficits (Semmler *et al.,* 2008), as well as early and sustained microglial activation, following peripheral LPS injection (Anderson *et al.,* 2015). Although sickness behavior disappears within days following LPS injection, long-term cognitive alterations may still be present, making mouse model of LPS-induced sepsis ideal for studying long-term cognitive deficits.

### 1.5. Determination and refinement of humane endpoints

Among animal models of sepsis, mouse models are widely used because of the large number of available genetically-engineered strains and the lower cost. Due to the nature of animal models of acute disease, high level of suffering and discomfort for experimental animals has been a concern in the implementation of the Three Rs (replacement, reduction and refinement; Russell, Burch and Hume, 1959). Accordingly, following the induction of acute disease, a method to identify animals at higher risk of death at an early time point to minimize unnecessary suffering is highly desired. To identify the endpoint by which experiments should be terminated without compromising data quality and interpretation, humane endpoints, a substitute for more severe experimental outcome such as spontaneous death, were introduced and applied in various animal models of disease.

However, among studies that reported the use of humane endpoints in animal models of disease, a high intra- and inter-model variance in endpoint determination and application, and a lack of performance metrics in the evaluation of humane endpoints have been observed (Mei *et al.,* 2019). Given the varying nature of animal models, unstandardized protocols in inspection and heterogeneity of previously used humane endpoints, accurate determination of humane endpoints remains challenging. Therefore, alternative methods to determine humane endpoint are needed to improve interpretability and reproducibility of research with animal models of sepsis, and to better meet the requirements of animal welfare.

### 1.6. Behavioral testing using a fully automated 8-arm radial arm maze

In experimental research, mazes are used to determine an animal's cognitive characteristics such as memory and attention. The radial arm maze (RAM) has been widely used in the study of spatial learning and memory since first described in 1976 (Olton and Samuelson, 1976). The RAM generally consists of a central platform from which 4-8 arms radiate outwards, each could be reward-baited with food, sugar pellets or water. Pre-experimental habituation is required to familiarize animals with the setup, during which an animal is placed on the central platform of the RAM for freely exploring the RAM and familiarizing with rewards in the baited radial arms. During experiments with the RAM, a reward is placed by the end of each radial arm (working memory paradigm) or selected arms (combined working/reference memory paradigm), and animals are required to remember the location of baited and visited arms to optimally obtain rewards.

In experiments with RAM, animals are usually food or water-restricted for prolonged periods of time to achieve sufficient motivation, which causes reduced body weight and altered appetite and may affect an animal's foraging strategy (Vorhees and Williams, 2014). A number of studies have also suggested learning outcomes could be affected by the level of motivation and food or water restriction (Pesic *et al.,* 2010; Johansson *et al.,* 2008; Vorhees and Williams, 2014) and revealed that water- or food-deprived animals may have different response patterns to the behavioral task (Reberg, Mann and Innis, 1977). Food or water restriction may lead to different motivation levels and learning outcomes, therefore potentially reduce reproducibility and interpretability of experimental results.

According to published studies, even minimal handling can induce stress that may create changes in animal behavior (Schellinck, Cyr and Brown, 2010; Hurst and West, 2010). Therefore, experiments with RAM which require a substantial degree of manual handling can have effects on stress responses that influence learning outcomes. As home-cage testing eliminates the need for handling (Schellinck, Cyr and Brown, 2010), it could lead to a promising alternative in behavioral testing with RAM. Although attempts have been made to automatize the RAM, the objectives were to ensure higher quality of data collection but not to fully meet the requirements of animal welfare, as food or water restriction was still broadly used. Therefore, to optimize interpretability and reproducibility of experimental results, minimized food and/or water restriction, reduced handling and home-cage testing are desirable.

### 1.7. Aim of the study

Microglial activation and phagocytosis may contribute to neuronal and synaptic loss during LPS-induced neuroinflammation. Following this assumption, we explored the role of deficiency of phagocytic pathways in post-sepsis behavior and cognitive alterations. The mouse model of experimental sepsis was established with intraperitoneal LPS injections in female homozygous wildtype C57BL/6 mice and female homozygous knockout and wildtype mice that are deficient for Mertk, Cd11b and Mfge8. To understand long-term cognitive alterations in LPS-injected mice, cognitive tests were performed following a one-month recovery period. For minimized handling-induced stress and unnecessary suffering, we also evaluated the use of automated behavioral testing systems and machine learning-

determined humane endpoints in behavioral experiments and sickness behavior monitoring, respectively. Following that, in the present study, three sub-projects have been defined:

1. Assessment of physiological and cognitive alterations in the mouse model of sepsis, and the effect of phagocytic deficiency or inhibition of phagocytic signaling pathways on cognitive functions;
2. Refinement of humane endpoints in animal models of acute disease; and
3. Automatization of experiments with 8-arm radial maze.

# 2. Materials and methods

## 2.1. Ethical statement

Animal use in all experiments was approved by the Landesamt für Gesundheit und Soziales (LaGeSo), Berlin. Experiments were conducted in accordance with the German animal protection law and local animal welfare guidelines.

## 2.2. Methods to prevent bias

For the sepsis model, animals were randomized for treatment, measurement modality, and survival times using the Research Randomizer tool (https://www.randomizer.org). For the stroke model, animals were randomized for treatment group using the GraphPad calculator tool (https://www.graphpad.com/quickcalcs/randomize1.cfm).

To minimize experimenter bias, randomization was conducted by a researcher who was not involved in injections, treatments, data acquisition or analysis. Information on strain, genotype and treatment group assignment was concealed from the experimenter until the end of the study.

## 2.3. Sub-project 1: Assessment of physiological and cognitive alterations in the mouse model of sepsis, and the effect of phagocytic deficiency or inhibition of phagocytic signaling pathways on cognitive functions

### 2.3.1. Animals

Two-month-old female mice (n = 435; Table 2.1) were used and housed individually in custom-made polycarbonate cages (dimensions: 20 × 20 × 30 cm) for video recording or type III polycarbonate cages, from 2 days before treatments until the 4th days after treatments before transported back to homecages. Food and water could be accessed *ad libitum*. Room temperature was maintained at 23.0 ± 1.0 °C with a humidity of 55-65%. A 12:12 h light/dark cycle with lights on at 8:00 was used (dark phase: 8:00-20:00; light phase: 20:00-8:00). Wood chips were used as bedding. Mice were randomly assigned to treatment groups and were used in experiments at the age of 8-10 weeks.

### 2.3.2. Drug administration

Animals were treated with LPS (from Salmonella enterica serotype, Sigma-Aldrich St. Louis, USA) at a dose of 1.5 mg/kg or physiological phosphate-buffered saline (PBS) solution through intraperitoneal administration on two consecutive days at 8:00 with a volume of 10 µl/g. To investigate the effect of inhibition of phagocytic signaling pathways, Cilengitide (Cilengitide©, Selleckchem Inc.), a vitronectin receptor antagonist, or inactive control peptide cRGD (Peptide Special Laboratories, GmbH), or PBS (10 ml/kg), was administered intraperitoneally for 7 days following LPS injections. Cilengitide or control peptide at a dose of 30 mg/kg was injected 6 hours after the 1st LPS injection, followed by daily injections at a dose of 10 mg/kg.

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

**Table 2.1:** Strain and origin of animals used in sub-project 1.

### 2.3.3. Temperature transponder implant

For temperature acquisition using radio-frequency identification (RFID) technology, passive RFID transponders were implanted subcutaneously. Temperature transponders (dimension: 2 × 14 mm; model: IPTT-300 transponders; BioMedic Data Systems, Seaford, USA) were programmed with individual identification numbers, loaded in a needle applicator device, and sterilized prior to implantation. Three weeks before the first LPS or saline injection, temperature transponders were implanted subcutaneously in the region between the scapulae. Anesthesia was induced with 2% isoflurane delivered in 100% oxygen for < 45 s before the implantation and injected once with meloxicam (1 mg/kg; Sigma-Aldrich, USA) for analgesia. Mice were observed for up to 48 hours for signs of complications and transponders were checked weekly for presence and functionality before the experiments.

### 2.3.4. Sickness behavior monitoring

**a.** Timeline

Baseline body temperature, weight and sickness severity score were obtained at 8:00 on the day of the 1st injection. During the two injection days and the two days following the 2nd injection, animals were inspected from 8:00 to 20:00. Body temperatures and sickness score

were obtained every 1.5 hours on each injection day. Body temperatures and sickness score were obtained every 6 hours for two days following the 2nd injection, and once per day at 8:00 from post-injection day 3. Body weight was obtained every 6 hours during the two injection days and the first two days following the 2nd injection, then once per day at 8:00.

**b.** Temperature acquisition

Core temperature was measured with a non-contact handheld transponder reader (DAS-7008/9; BioMedic Data Systems, Seaford, USA) held above the animal's shoulder region. Two non-contact infrared thermometers (model 1: Braun No touch – NTF3000; Braun, Kronberg, Germany; model 2: Aponorm Contact Free 3; WEPA Aponorm, Hillscheid, Germany) were used in surface temperature measurement in the perianal region.

**c.** Sickness severity scoring

Sickness severity was scored on a scale of 0 - 5 (0: normal level of activity; 5: maximum sickness severity; Mei *et al.,* 2018).

**d.** Body weight measurements

A bench scale was used (PCB 1000-1, KERN & SOHN GmbH, Balingen, Germany). Mice were weighed when both their body and tail were in a plastic box placed on the scale. Body weights were taken when a stable reading has been displayed.

**e.** Video recording and nest scoring

A GigE monochrome camera (DMK 33GR0134, The Imaging Source Europe GmbH, Bremen, Germany) was used. To reduce interference, body temperature and weight were measured with a 6-hour interval (at 8:00, 14:00 and 20:00). Sickness severity was scored every 1.5 hours by observing the animals through the transparent lid of the cage. Ambulation distance and percentage of activity (time of ambulation divided by total time) were computed with automated tracking software (Viewer, Biobserve GmbH, Bonn, Germany). Scoring criteria of nest building behavior were adapted from Deacon (2006). Nest building was evaluated by two experimenters through the video daily at 8:00. The average score was used in the analysis.

**f.** Humane endpoint

Upon reaching a sickness severity score > 4 once, or a score = 4 twice within 2 hours, an animal was removed from the cage and immediately euthanized by cervical dislocation.

### 2.3.5. Experiments

Experiments started a month following LPS or saline injections and were conducted with 2 IntelliCages (NewBehavior/TSE-Systems GmbH, Bad Homburg, Germany), in which mice were housed in original groups. IntelliCage is a homecage-based monitoring system that allows automated recording and assessment of spontaneous and cognitive behavior in transponder-implanted mice in a social setting without human interference (Lipp *et al.*, 2005). An IntelliCage consisted four operant learning corners in each two water bottles were placed, with the nozzle behind a motorized door for programmable access to water. Food was accessible *ad libitum*.

The experiment consisted of habituation (free exploration, nosepoke adaptation and drinking session adaptation) and cognitive testing (Figure 2.1). Free exploration allowed *ad libitum* access to water bottles. During nosepoke adaptation, animals needed to use a nosepoke to open the motorized door and access water bottles. In drinking session adaptation, doors could be opened with nosepoke during drinking sessions (10:00-12:00 and 18:00-20:00) and remain closed during non-drinking sessions (20:00-10:00 and 12:00-18:00). Upon completion of the 3 habituation paradigms, 4 paradigms were used to assess cognitive aspects including spatial learning (place learning and place reversal), reward-driven behavior (sucrose preference) and avoidance learning and retention (avoidance conditioning).

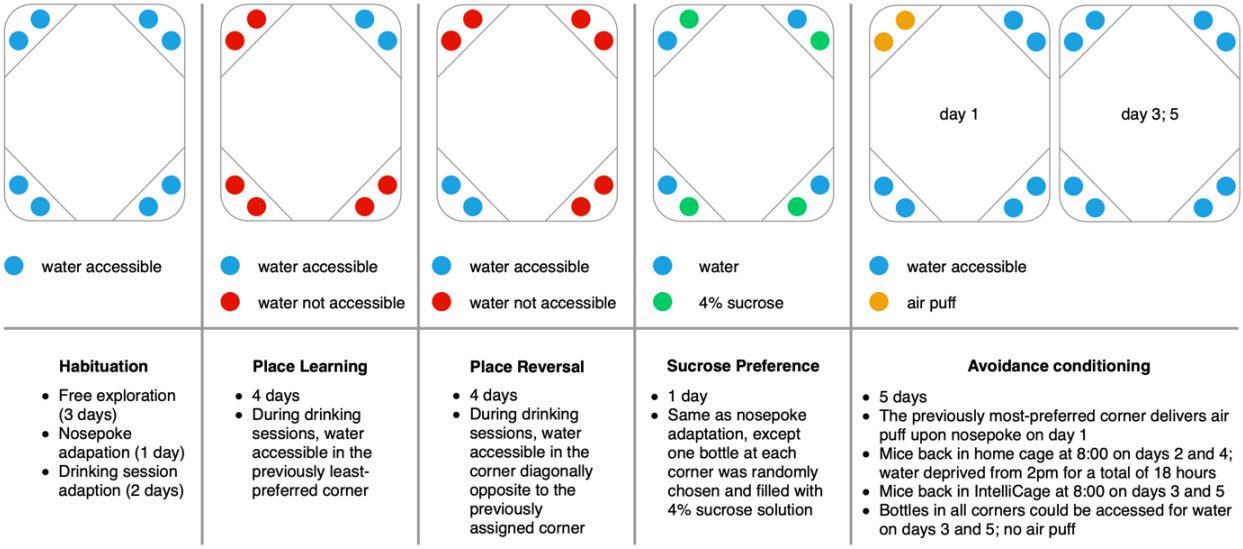


**Figure 2.1:** Habituation and cognitive testing paradigms in IntelliCage.

### 2.3.6. Statistical analysis

**a.** Sickness behavior monitoring

Results are expressed as mean (SD). Data processing and statistical analysis was performed using SPSS Statistics version 24 (IBM Corp., USA) and Python 2.7.10 (Python Software Foundation).

**b.** Behavioral testing

Data was analyzed with PyMice (Dzik *et al.,* 2018), Python 2.7.10 (Python Software Foundation) and SPSS Statistics version 25 (IBM Corp., USA). Shapiro-Wilk test was used to examine the normality of data. For normally distributed samples, T test was used. Mann Whitney U test was used when the samples deviated from normal distribution. One-way ANOVA with Tukey's procedure to correct for multiple comparisons was used for comparisons across three groups. For samples that deviated from normal distribution, Kruskal-Wallis test was used. Results were reported as mean ± standard error of mean (SEM). P-values < 0.05 were considered statistically significant.

## 2.4. Sub-project 2: Refinement of humane endpoints in animal models of acute disease

### 2.4.1. Animals

All data analyzed were sourced from previous results and no animals were used in this sub-project. Data from 922 animals were included. For the stroke model, 487 animals (Table 2.2) were used and randomly assigned to treatment groups at the age of 8 - 12 weeks. Mice were housed in groups of up to 12 at a room temperature of 22 ± 2°C with a humidity of 55 ± 10%. A 12:12h light/dark cycle with lights on at 7:00 was used (dark phase: 7:00-19:00; light phase: 19:00-7:00). Aspen woodchips were used as bedding. For the sepsis model, see Section 2.3.1.

| 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) | |

**Table 2.2:** Strain and origin of animals used in the stroke model. m: male; f: female.

### 2.4.2. Treatments

In the 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. For the sepsis model, see Section 2.3.2.

### 2.4.3. Physiological monitoring

In the stroke model, body weight and a modified version of the DeSimoni neuroscore were obtained. Core body temperature was quantified non-invasively using subcutaneous RFID transponders as described in Donath *et al.* (2016). For the sepsis model, see Section 2.3.4.a-e.

### 2.4.4. Humane endpoint

In the stroke model, animals were euthanized by cervical dislocation upon reaching a score of 2 of the 2$^{nd}$ criteria, or a score of 3 or 4 of the 3$^{rd}$ -12$^{th}$ criteria in the modified DeSimoni neuroscore (Donath *et al.*, 2016). In addition, animals were euthanized when a loss > 20% baseline body weight occurred or the following signs were observed: 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. For the sepsis model, see Section 2.3.4.f.

## 2.5. Sub-project 3: Automatization of experiments with 8-arm radial maze

### 2.5.1. Animals

Twelve female homozygous wildtype C57BL/6J mice were derived from Charles River at 6-8 weeks of age and housed locally. Two animals were excluded prior to the beginning of the experiment due to a dislodged/malfunctioning RFID transponder (n = 1) or death from a natural cause (n = 1). Mice were housed under standard animal laboratory conditions and transponder-implanted as in sub-project 1. Ten animals were subjected to the working memory paradigm and combined working/reference memory paradigm at the age of 18 months. Of those, four randomly selected animals had been subjected to the working

memory paradigm once at the age of 9 months to ensure general functionality of the automated 8-arm RAM during its development.

### 2.5.2. Experimental apparatus

**a.** Automated 8-arm RAM

The custom-made automated RAM (PhenoSys GmbH, Berlin, Germany) was constructed of eight transparent plastic tubes (diameter: 4.8 cm; length: 50 cm) radiating outwards from an octagonal central platform (span: 23.5 cm; side length: 9.7 cm; height: 12 cm; Figure 2.2, a). A plastic panel was placed on the central platform to allow for additional tactile cues and its orientation remained unchanged during the experiment (Figure 2.2, b). The central platform was covered with a transparent lid to prevent animals from escaping.

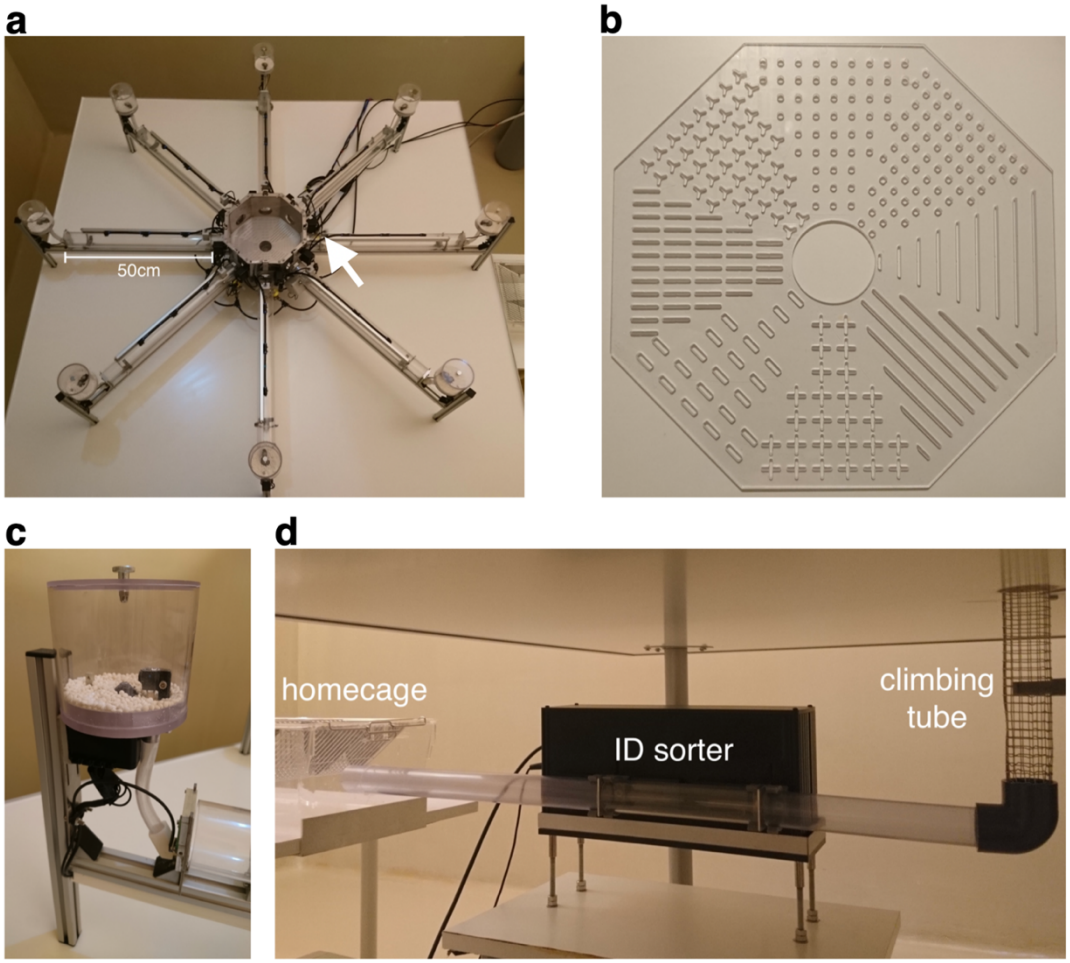


**Figure 2.2:** Setup of the automated RAM. (a). Top view of the main body of the automated RAM. Arrow, position of the infrared sensors. (b). Plastic panel of tactile cues to be placed on the central platform of the RAM. (c). Automated sugar pellet dispenser. (d). RFID transponder-based ID sorter and climbing wire connecting the homecage and the RAM.

An automated sugar pellet dispenser was located at the end of each arm (Figure 2.2, c). Upon first entry to a baited arm, an animal was detected by the infrared sensor, which triggered one sugar pellet (Purified Rodent Tablets 5TUL, Test Diet, Richmond, IN, USA) to be dispensed. No sugar pellet was dispensed in an un-baited arm. When all baited arms have been visited and an animal has returned to the central platform, mechanical doors at the entrance of radial arms are closed to prevent further entries.

To allow animals to enter the automated RAM, a sorter (IDsorter, PhenoSys GmbH, Berlin, Germany; Figure 2.2, d) was placed between the homecage and the RAM. The sorter read an animal's RFID transponder and granted 24/7 access to the maze as long as an animal's ID was registered in the experiment's configuration file and there was no other animal inside the maze. The homecage and sorting device were both placed below a plastic lab table (dimensions: 160 x 160 cm); the RAM was placed on top (81 cm above floor level). A wire-mesh climbing tube runs through a hole located in the middle of the tabletop, connecting the sorter with the central platform of the RAM, allowing animals to climb up to the RAM (Figure 2.2, d). In addition to the intra-maze tactile cues, extra-maze visual cues were used during the experiment.

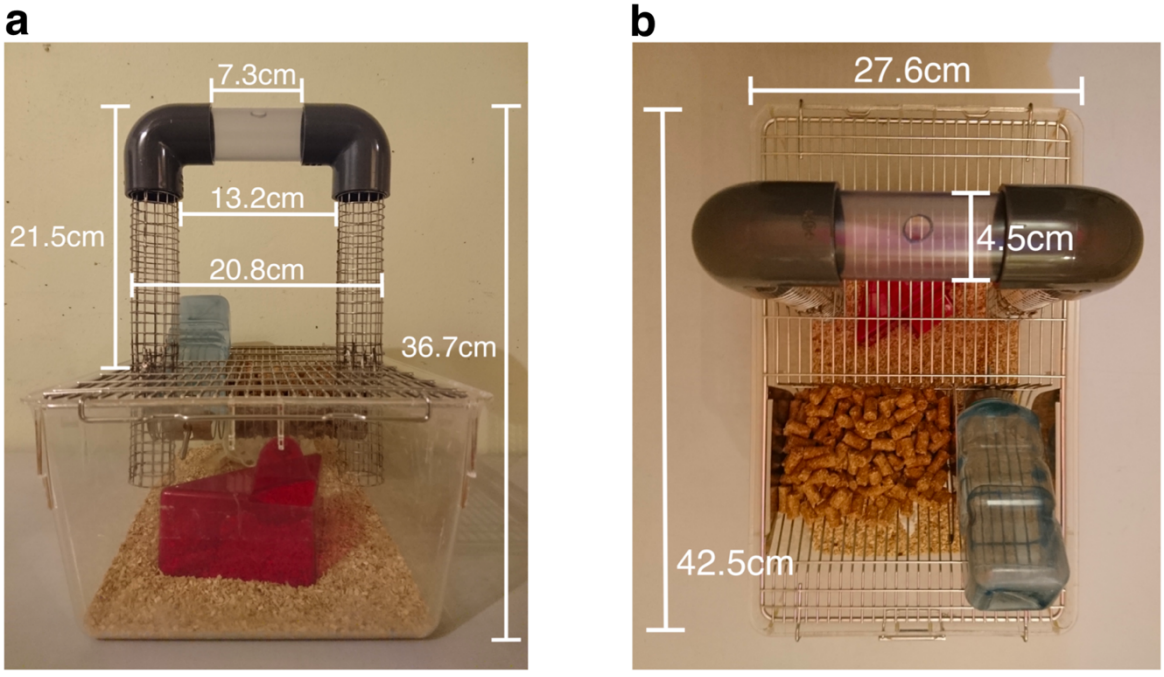


**Figure 2.3:** Custom-made cage for habituation. (a). Front view. (b). Top view.

**b.** Custom-made cage for habituation

For familiarization, mice were house in a polycarbonate cage (dimensions: 42.5 x 27.6 x 15.3 cm; 1290D Eurostandard Type III, Techniplast, Italy) with a custom-made lid (Figure 2.3). A borehole (diameter: 1cm) on the upper rim of the plastic tube allowed for sugar pellets to be placed for habituation purposes.

**c.** Software

The software controlling the RAM (PhenoSys GmbH, Berlin, Germany) allows data on behavioral event, component that registered the event (e.g., dispenser, door, infrared sensor, RFID reader) and duration of the event to be recorded. Data was exported for offline analysis after each experiment. Mechanical components of the RAM could be controlled and inspected through the software.

### 2.5.3. Experiments

Experiments with the RAM included three phases: habituation (3 days), working memory paradigm (7 days), and combined working/reference memory paradigm (5 days). During working memory paradigm and combined working/reference memory paradigm, animals could voluntarily access the automated RAM at any time.

- Habituation: Animals were housed in the custom-made cage (Figure 2.3) to be familiarized with sugar pellets, which served as positive rewards during RAM experiments, and the wire-mesh climbing tunnel. Sugar pellets were placed inside the climbing piece of the custom-made cage daily at 9:00. Following habituation, mice were manually transferred to their homecage and were housed for the remainder of the experiment.
- Working memory paradigm: An animal's first entry into each radial arm triggered one sugar pellet to be dispensed at the end of the arm. The animal was able to collect a maximum of 8 sugar pellets before the end of each session.
- Combined working/reference memory paradigm: Sugar pellets were dispensed upon the first entry into the 4 randomly chosen baited arms. The animal was able to collect a maximum of 4 sugar pellets before the end of each session.

### 2.5.4. Exclusion criteria

Data entries displaying one or more of the following errors were excluded from further analysis: (a). Malfunctioning pellet dispenser that did not dispense a sugar pellet upon an

animal's first entry into a baited arm, (b). Incorrect registration of re-entries into visited arms by the infrared sensor, and (c). Incorrect registration of a session's duration or paradigm-related information.

### 2.5.5. Statistical analysis

Data are shown in mean ± standard error of mean (SEM). A Python script (Python 2.7.15, Python Software Foundation) was used to automatically process raw data files and calculate all parameter values. Processed data was analyzed and plotted with GraphPad Prism software version 8.0.0 (GraphPad Prism Software, San Diego, CA). To assess the effect of time, one-way repeated measures ANOVA was used and corrected for sphericity using the Greenhouse-Geisser correction when the assumption of sphericity was violated. P-values $< 0.05$ were considered statistically significant.

# 3. Results

## 3.1. Sub-project 1: Assessment of physiological and cognitive alterations in the mouse model of sepsis, and the effect of phagocytic deficiency or inhibition of phagocytic signaling pathways on cognitive functions

### 3.1.1. Mortality

Two-hundred and fifty-four out of 284 LPS-treated animals (89.4%) reached the pre-defined experimental endpoint. Eighteen of the 30 dead animals were found dead and 12 were euthanized upon reaching pre-defined humane endpoint. Death occurred 24-192 hours after the 1st injection (mean = 60.5 (35.1) hours). All saline-injected mice reached the pre-defined experimental endpoint. To facilitate comparison, animals were divided into three groups:

1. Control (n = 151): treated with saline; reached the planned experimental endpoint;
2. Survivor (n = 254): treated with LPS; reached the planned experimental endpoint;
3. Non-survivor (n = 30): spontaneously died or were euthanized during sickness behavior monitoring.

### 3.1.2. Post-treatment physiological changes in saline- and LPS-injected animals

**a.** Core and surface temperatures

No significant decrease in core and surface temperatures from baseline was observed in control animals (Figure 3.1, a, b). Survivor and non-survivor groups showed decreased core and surface body temperatures during the days of LPS injections. In the survivor group, lowest core temperature of both injection days was observed 12 hours following injection (day 1: 34.1 (2.3) °C; day 2: 35.3 (2.1) °C). In the non-survivor group, lowest core temperature was observed 12 (31.2 (2.3) °C) and 7.5 (25.4 (4.5) °C) hours following LPS injection on injection day 1 and 2.

In the survivor group, lowest surface temperature of both injection days was observed 12 hours following injection (day 1: 27.8 (1.9) °C; day 2: 28.4 (1.9) °C). Surface body temperature of the survivor group returned to baseline level within 96 hours following the 1st injection. Surface temperature of non-survivor animals was the lowest 12 and 9 hours following injection on injection day 1 and 2 (25.7 (1.8) °C and 23.7 (1.8) °C).

**b.** Sickness severity score

Sickness severity score of control animals remained unchanged (Figure 3.1, c). Sickness severity score of the survivor group peaked 12 hours following injections on both day 1 and 2 (1.2 (0.9)) and (1.4 (0.9)) and returned to baseline level within 96 hours following the 1st injection. Sickness score of non-survivors was at the daily maximum 12 hours after each injection (day1: 2.1 (0.7); day 2: 3.2 (0.7)).

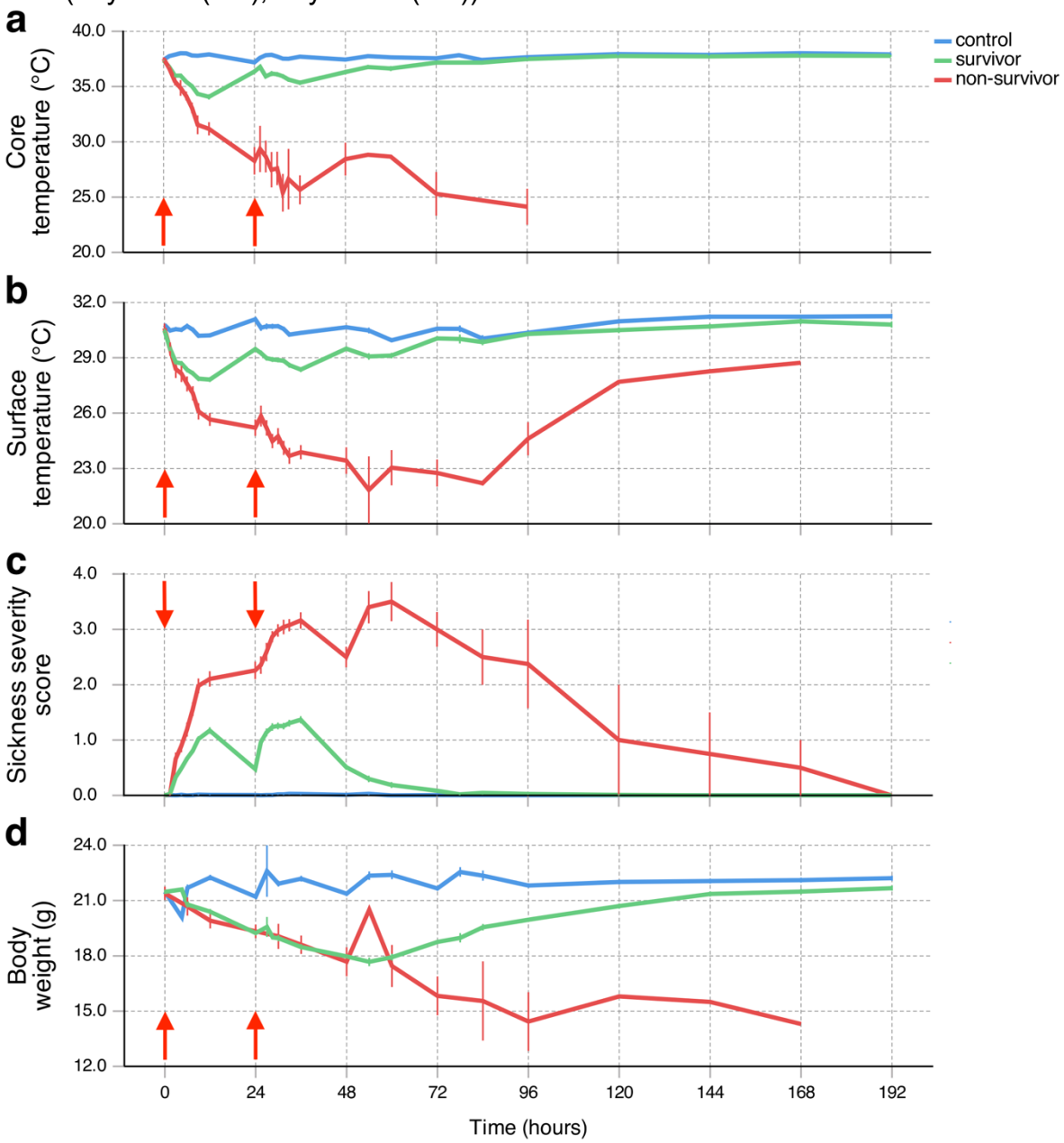


**Figure 3.1:** Physiological changes in LPS- and saline-injected animals during sickness behavior monitoring. (a). Core temperature. (b). Surface temperature. (c). Sickness severity score. (d). Body weight. Red arrow, time of LPS/saline injections. Data are presented as mean ± SEM.

**c.** Body weight

No weight changes were observed in the control group (Figure 3.1, d). Body weight of the survivor group reached its minimum 54 hours following the 1st injection (17.7 (1.5) g) and returned to baseline level 168 hours following the 1st injection. Lowest weight of non-survivor animals was observed 96 hours after the 1st injection (14.4 (2.8) g).

**d.** Activity level and nest building behavior

All groups displayed comparable baseline activity and novelty-induced high ambulation level (Figure 3.2). A day-night rhythm was observed in control animals. Sustained decrease in the level of activity was observed from the 1st injection day until the end of video recording in survivor and non-survivor groups. A rebound in ambulation distance and percentage of activity was observed in the survivor group 48 hours after the 1st injection, but not in the non-survivor group.

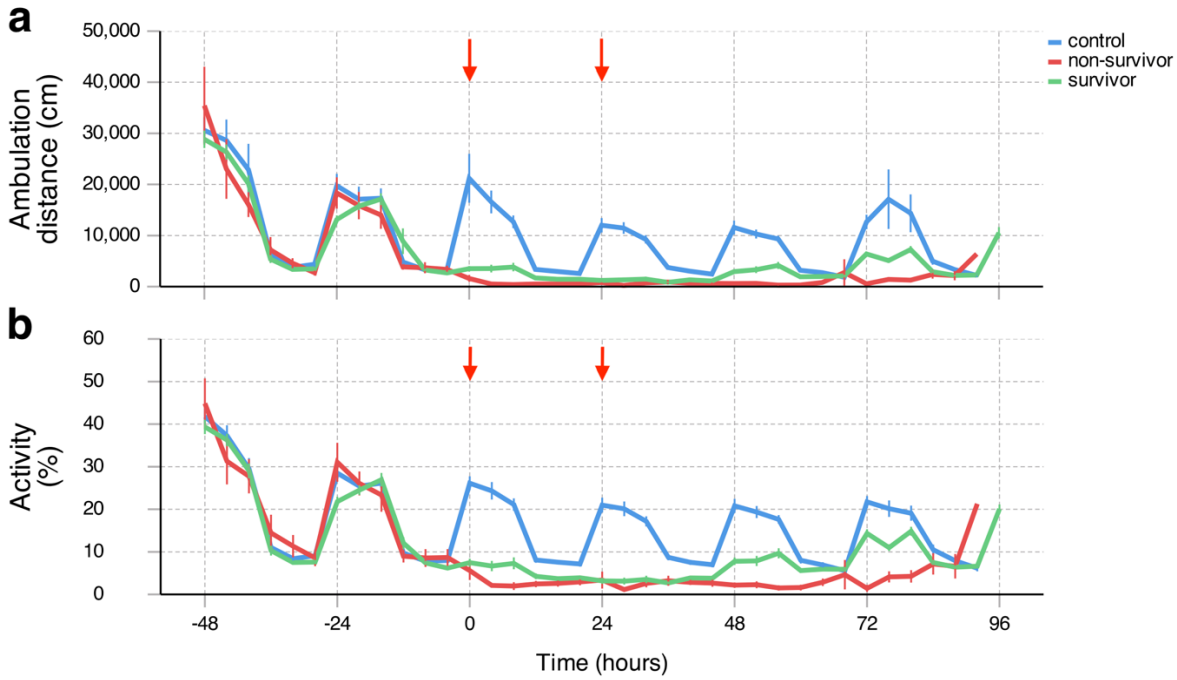


**Figure 3.2:** Activity level of LPS- and saline-injected animals during sickness behavior monitoring. (a). Distance of ambulation. (b). Percentage of activity. Red arrow, time of LPS/saline injections. Data are plotted in 4-hour bins and presented as mean ± SEM.

Nest building was comparable across groups at baseline. Both non-survivor group and survivor group displayed a decrease in nest building score after LPS injections, followed by a slow recovery in the survivor group and a further decrease in the non-survivor group (Figure 3.3)

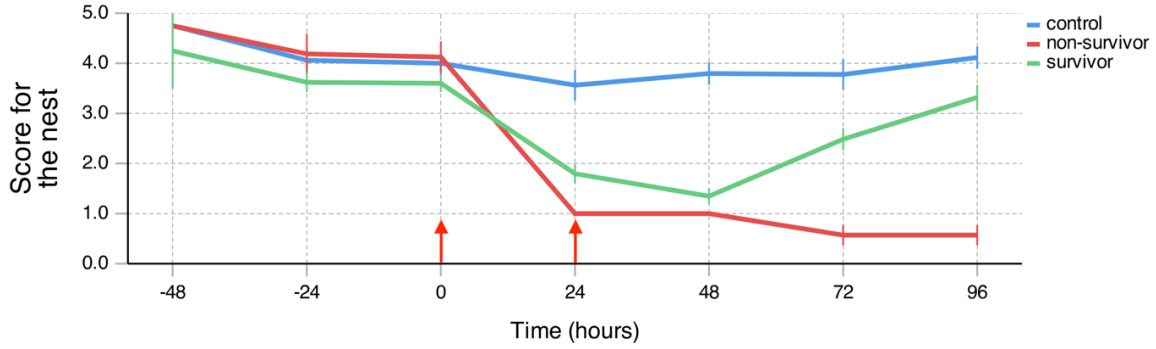


**Figure 3.3:** Nest building score of LPS- and saline-injected wildtype mice during habituation. Red arrow, time of LPS/saline injections. Data are presented as mean ± SEM.

### 3.1.3. Cognitive alterations in LPS-injected mice and the effect of inhibition of microglial phagocytosis

**a.** Saline- and LPS-injected wildtype mice

**Habituation**: During free exploration, both groups displayed higher level of activity in the dark phase (Figure 3.4, a-c). LPS-injected mice had more visits, nosepokes and licks (p = 0.003, 0.007 and 0.001, respectively) during the dark phase of the 1st day of free exploration. Saline- and LPS-injected mice required comparable amount of time to achieve 50 nosepokes in nosepoke adaptation (p = 0.054; Figure 3.4, d).

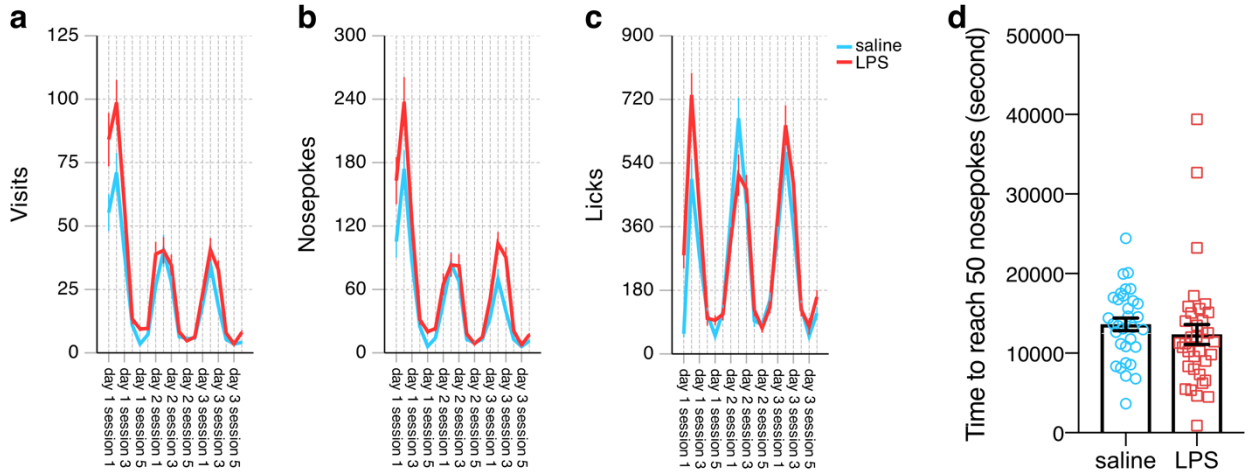


**Figure 3.4:** Activity of LPS- and saline-injected wildtype mice during habituation. (a-c). Free exploration. (d). Nosepoke adaptation. Data are plotted in 4-hour bins for free exploration and presented as mean ± SEM.

**Cognitive tests:**

Place learning: Saline-injected mice had higher percentage of nosepokes at the water-rewarded corner ("correct nosepokes") in the 4-day place learning paradigm (p = 0.009;

Figure 3.5, a), and in the first 100 ($p < 0.001$; Figure 3.5, c) and 150 nosepokes ($p = 0.03$; Figure 3.5, d).

Place reversal: LPS- and saline-injected mice had comparable percentage of correct nosepokes in the 4-day place reversal paradigm ($p = 0.197$; Figure 3.5, e). Saline-injected mice had higher percentage of correct nosepokes in the first 50, 100 and 150 nosepokes ($p = 0.002$, 0.004 and 0.003; Figure 3.5, f-h).

Sucrose preference: Increase in preference towards sucrose-rewarded bottles was comparable across treatments ($p = 0.087$; Figure 3.6, a).

Avoidance conditioning: Avoidance towards the previously air-puffing corner was comparable across treatments during training ($p = 0.158$; Figure 3.6, b). Higher avoidance retention was observed in saline-injected mice during test and re-test ($p = 0.017$ and 0.001).

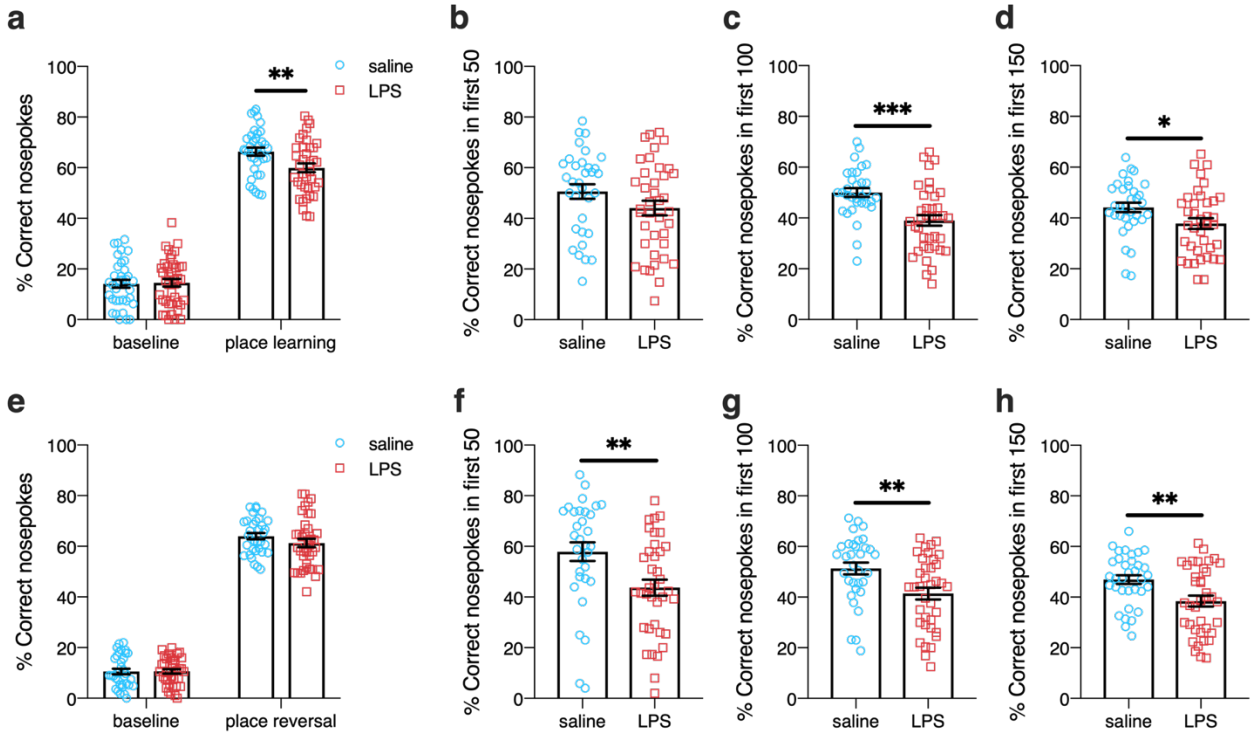


**Figure 3.5:** Spatial learning of LPS- and saline-injected wildtype mice. (a). Percentage of correct nosepokes in the 4-day place learning paradigm. (b-d). Percentage of correct nosepokes in the first 50, 100 and 150 nosepokes of place learning. (e). Percentage of correct nosepokes in the 4-day place reversal paradigm. (f-h). Percentage of correct nosepokes in the first 50, 100 and 150 nosepokes of place reversal. Data are presented as mean ± SEM; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

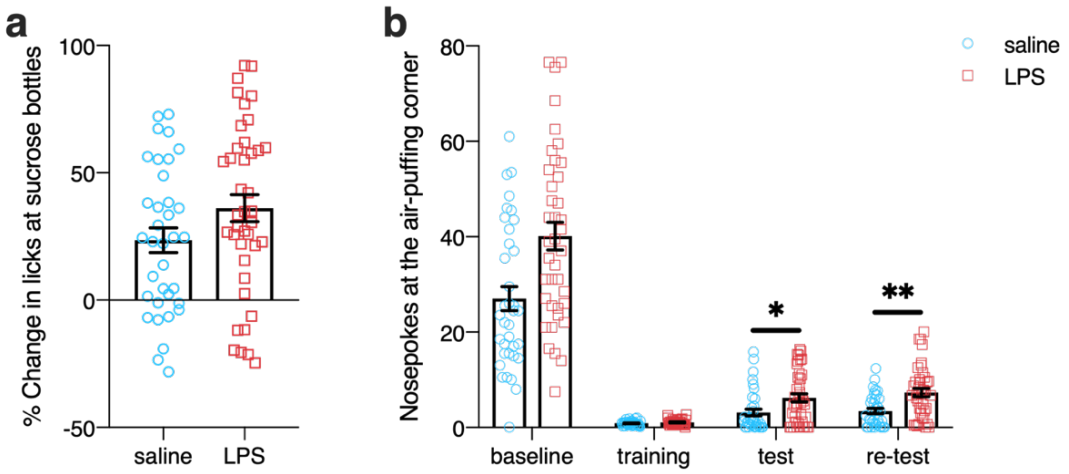


**Figure 3.6:** Reward-driven behavior, avoidance conditioning and retention of LPS- and saline-injected wildtype mice. (a). Change in preference towards sucrose-rewarded bottles. (b). Number of nosepokes at the air-puffing corner at baseline, during training (air-puffing phase), test and re-test (retention phase). Data are presented as mean ± SEM; *p < 0.05, **p < 0.01.

**b.** LPS-injected wildtype and knockout mice

**Habituation:**

- Mertk: During free exploration, animals displayed higher activity level in the dark phase (Figure 3.7, a-c). A decrease in behavioral events but licks was observed in both groups after the 1st day of free exploration. Wildtype mice had more visits and nosepokes than knockout mice in the dark phase of day 1 of free exploration (p = 0.003 and 0.045).
- Cd11b: For free exploration, data were not available due to mechanical errors. Knockout mice required less time to achieve 50 nosepoke (p = 0.031; Figure 3.7, d).
- Mfge8: During free exploration, animals showed higher activity level in the dark phase (Figure 3.7, a-c). The number of visits, nosepokes and licks was comparable across genotypes (p = 0.351, 0.098 and 0.635, respectively).

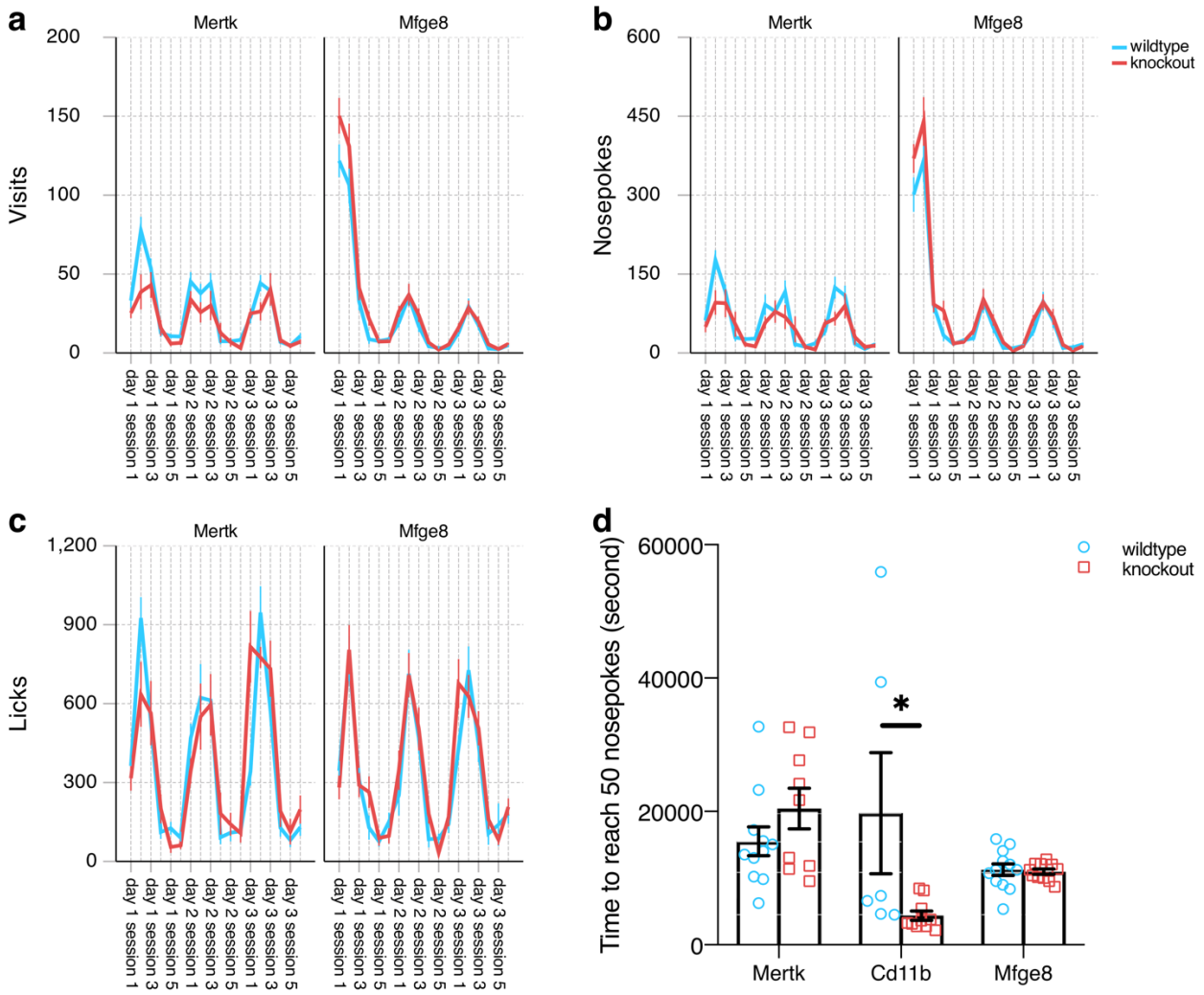


**Figure 3.7:** Activity of LPS-injected wildtype and knockout mice during habituation. (a-c). Number of visits, nosepokes and licks in free exploration, (d). Time required to reach 50 nosepokes during nosepoke adaptation. Data are plotted in 4-hour bins for free exploration and presented as mean ± SEM; *p < 0.05.

**Cognitive tests:**

Place learning:

- Mertk: LPS-injected wildtype and knockout mice had comparable percentage of correct nosepokes in the 4-day place learning paradigm (p = 0.222; Figure 3.8, a) and in the first 50, 100 and 150 nosepokes (p = 0.650, 0.636 and 0.522, respectively; Figure 3.8, b-d).
- Cd11b: Percentage of correct nosepokes during place learning and in the first 50, 100 and 150 nosepokes was not significantly different across genotypes (p = 0.732, 0.990, 0.665 and 0.711, respectively; Figure 3.8, a-d).

- Mfge8: In the 4-day place learning paradigm, wildtype mice showed higher percentage of correct nosepokes (p = 0.017). Wildtype and knockout mice displayed comparable percentage of correct nosepokes in the first 50, 100 and 150 nosepokes (p = 0.347, 0.412 and 0.434, respectively; Figure 3.8, b-d).

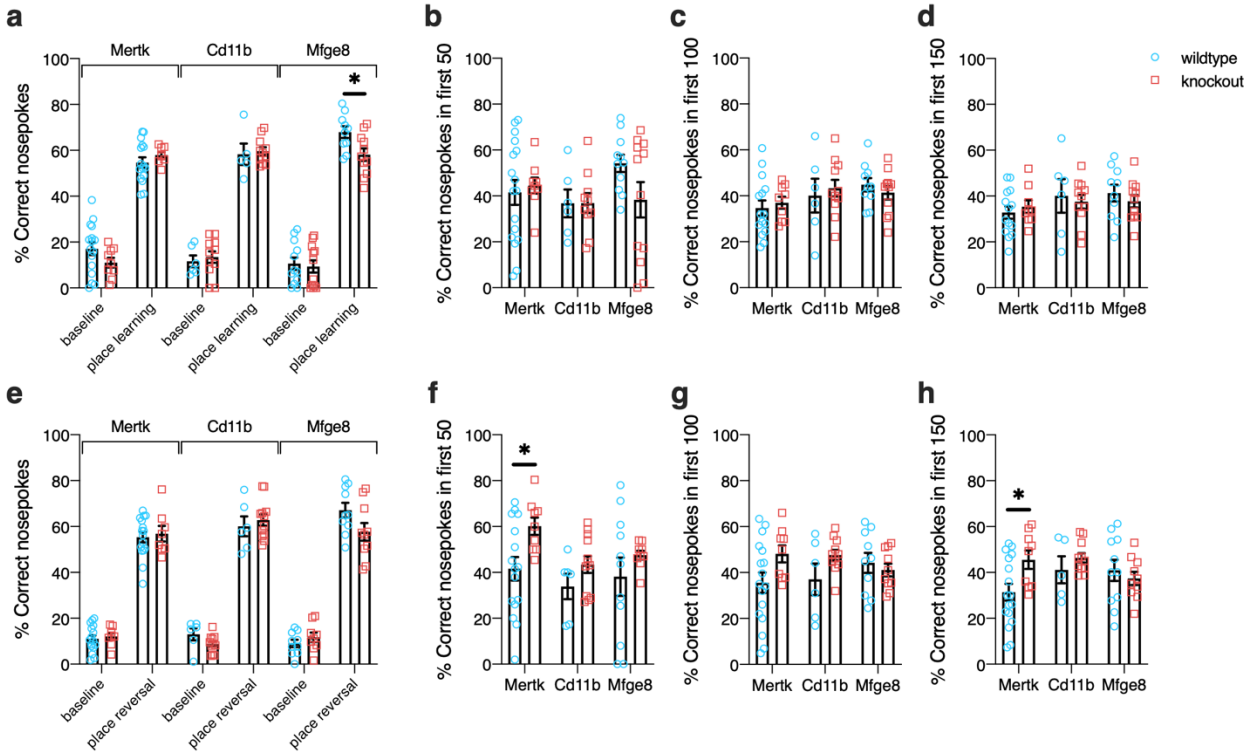


**Figure 3.8:** Spatial learning of LPS-injected wildtype and knockout mice. (a). Percentage of correct nosepokes in the 4-day place learning paradigm. (b-d). Percentage of correct nosepokes in the first 50, 100 and 150 nosepokes of place learning. (e). Percentage of correct nosepokes in the 4-day place reversal paradigm. (f-h). Percentage of correct nosepokes in the first 50, 100 and 150 nosepokes of place reversal. Data are presented as mean ± SEM; *p < 0.05.

Place reversal:

- Mertk: Percentage of nosepokes at the water-awarded corner was comparable across genotypes in the place reversal paradigm (p = 0.689; Figure 3.8, e). In the first 50 and 150 nosepokes, knockout mice showed higher percentage of correct nosepokes (p = 0.02 and 0.022; Figure 3.8, f, h).
- Cd11b: Percentage of correct nosepokes was comparable across genotypes in the place reversal paradigm and in the first 50, 100 and 150 nosepokes (p = 0.559, 0.162, 0.192 and 0.441, respectively; Figure 3.8, e-h).

- Mfge8: Percentage of correct nosepokes was comparable across genotypes in the place reversal paradigm and in the first 50, 100 and 150 nosepokes (p = 0.078, 0.296, 0.553 and 0.53; Figure 3.8, e-h).

Sucrose preference: The degree of increase in preference towards sucrose-rewarded bottles was comparable across genotypes in Mertk and Cd11b mice (p = 0.829 and 0.118; Figure 3.9, a). Increase in preference towards sucrose-rewarded bottles was higher in wildtype Mfge8 mice (p = 0.011; Figure 3.9, a).

Avoidance conditioning:

- Mertk: The two groups displayed comparable numbers of nosepokes at the air-puffing corner during training, test and re-test (p = 0.264, 0.787 and 0.698; Figure 3.9, b).
- Cd11b: Number of nosepokes at the air-puffing corner was comparable across genotypes during training, test and re-test (p = 0.057, 0.094 and 0.482; Figure 3.9, b).
- Mfge8: The two genotypes displayed comparable number of nosepokes at the air-puffing corner during training (p = 0.057). In test and re-test, knockout mice showed greater avoidance towards the air-puffing corner (p = 0.029 and 0.029; Figure 3.9, b).

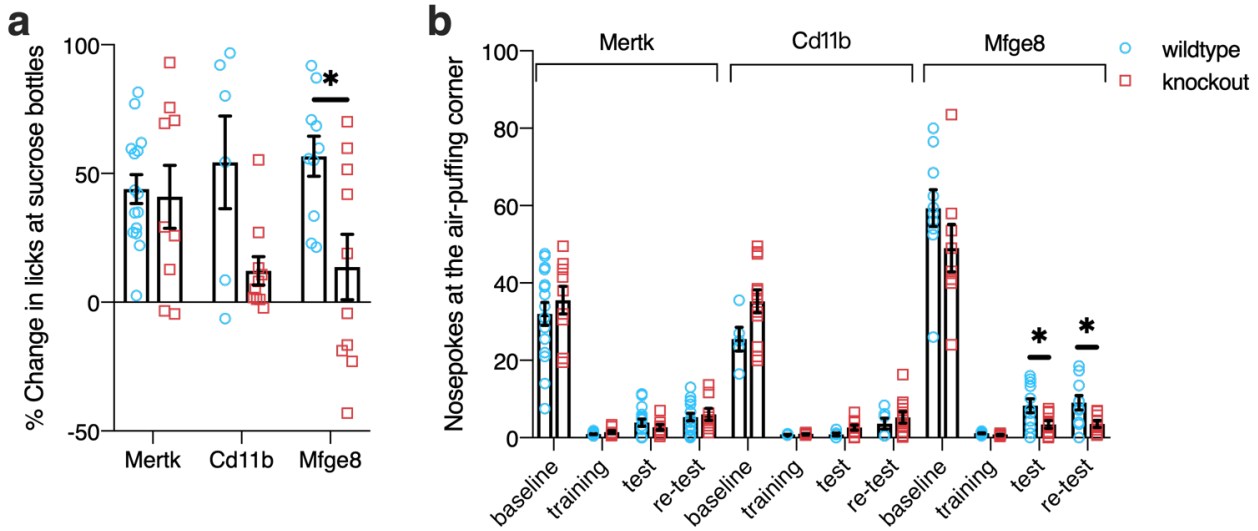


**Figure 3.9:** Reward-driven behavior, avoidance conditioning and retention of LPS-injected wildtype and knockout mice. (a). Change in preference towards sucrose-rewarded bottles. (b). Number of nosepokes at the air-puffing corner. Data are presented as mean ± SEM; *p < 0.05.

**c.** LPS-injected untreated and peptide-treated mice

**Habituation:** During free exploration, mice displayed higher activity level in the dark phase (Figure 3.10, a-c). Number of visits differed across treatment groups (p = 0.017). Peptide treatments led to a significantly higher number of licks (p = 0.006; Figure 3.10, c). Cilengitide- and cRGD-treated mice displayed more licks than untreated mice (p = 0.001 and 0.028) and had comparable licks (p = 0.281). During nosepoke adaptation, the time required to achieve 50 nosepokes was comparable across treatments (p = 0.127; Figure 3.10, d).

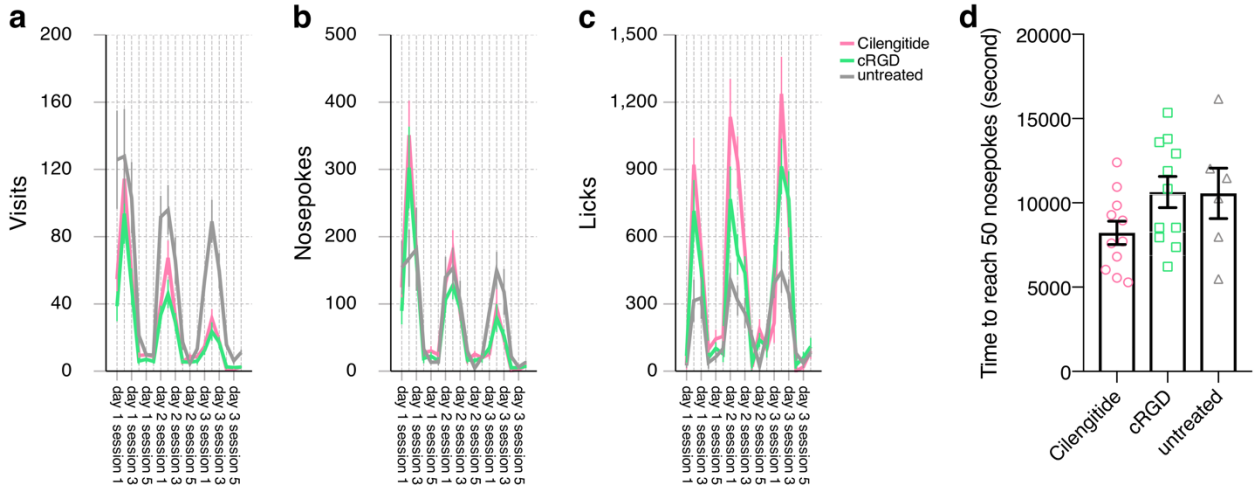


**Figure 3.10:** Activity of Cilengitide-treated, cRGD-treated and untreated LPS-injected wildtype mice during habituation. (a-c). Free exploration. (d). Nosepoke adaptation. Data are plotted in 4-hour bins for free exploration and presented as mean ± SEM.

**Cognitive tests:**

Place learning: Untreated mice displayed higher percentage of correct nosepokes than Cilengitide- or cRGD-treated mice (p = 0.026 and 0.009; Figure 3.11, a). Cilengitide-treated mice showed higher percentage of correct nosepokes in the first 50, 100 and 150 nosepokes than cRGD-treated mice (p = 0.001, 0.028 and 0.023; Figure 3.11, b-d).

Place reversal: Percentage of correct nosepokes was comparable across treatments in the 4-day place reversal paradigm (p = 0.073; Figure 3.11, e) and in the first 50, 100 and 150 nosepokes (p = 0.625, 0.687 and 0.312; Figure 3.11, f-h).

Sucrose preference: Increase in licks at the sucrose-rewarded bottles differed across treatments (p = 0.001). Cilengitide- and cRGD-treated mice displayed greater increase in preference towards sucrose-rewarded bottles than untreated mice (p < 0.001 and 0.011; Figure 3.12, a).

Avoidance conditioning: During training, test and re-test, avoidance towards the air-puffing corner was comparable across treatments (p = 0.435, 0.184 and 0.308, respectively; Figure 3.12, b).

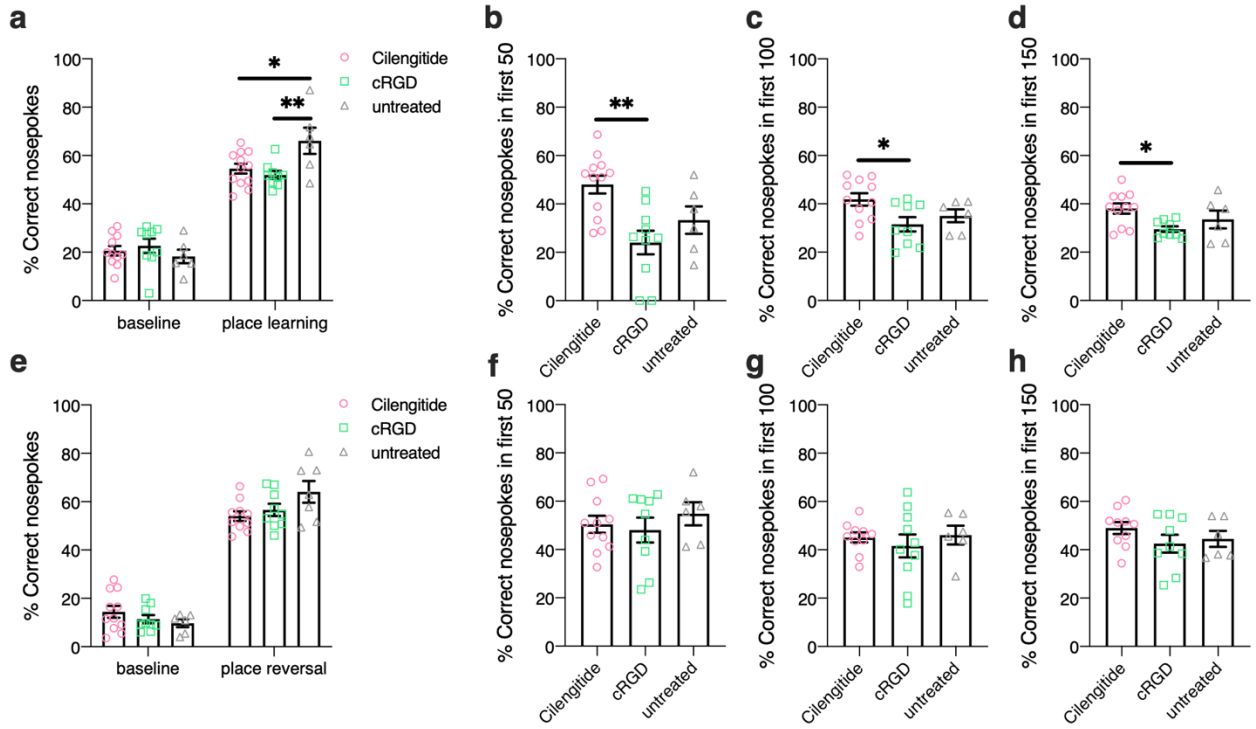


**Figure 3.11:** Spatial learning of Cilengitide-treated, cRGD-treated and untreated LPS-injected wildtype mice. (a). Percentage of correct nosepokes in the 4-day place learning paradigm. (b-d). Percentage of correct nosepokes in the first 50, 100 and 150 nosepokes of place learning. (e). Percentage of correct nosepokes in the 4-day place reversal paradigm. (f-h). Percentage of correct nosepokes in the first 50, 100 and 150 nosepokes of place reversal. Data are presented as mean ± SEM; *p < 0.05, **p < 0.01.

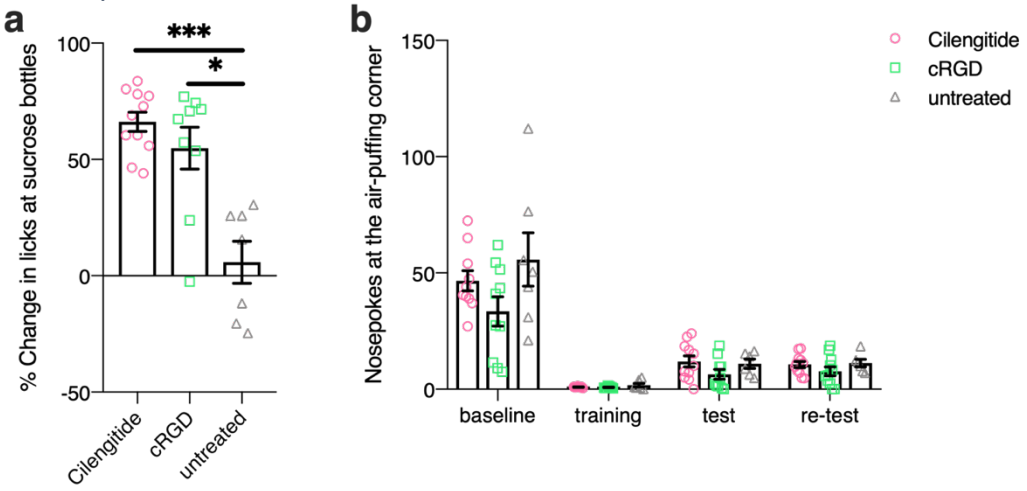


**Figure 3.12:** Reward-driven behavior, avoidance conditioning and retention of Cilengitide-treated, cRGD-treated and untreated LPS-injected wildtype mice. (a). Change in preference towards sucrose-rewarded bottles. (b). Number of nosepokes at the air-puffing corner. Data are presented as mean ± SEM; *p < 0.05, **p < 0.01, ***p < 0.001.

## 3.2. Sub-project 2: Refinement of humane endpoints in animal models of acute disease

### 3.2.1. Death prediction with core and surface temperatures

First, core and surface temperatures from transponder-implanted animals were used separately to train the prediction models to compare the accuracy of core and surface temperatures in the prediction of death (n = 160). Secondly, the best performing models were trained with data from all animals whose surface temperature measurements were available 36 hours following the 1st injection (n = 372) to test the effects of increased size of the training data on model performance.

At 36 hours after the 1st injection, death could be predicted with high accuracy both from core and surface temperatures (accuracy = 96.3% and 95.6% for core and surface temperatures, respectively; n = 160; number of dead animals = 13). Surface temperature data from 372 mice (number of dead animals: 28) led to further increase in accuracy to 96.5%.

### 3.2.2. Accuracy of death prediction

Death could be predicted with an accuracy of 93.2% (male) or 93.0% (female) in the model of stroke and 96.2% in the model of sepsis, with Gaussian Naïve Bayes (stroke model) or decision tree of depth 2 (sepsis model) (Table 3.1).

In the stroke model, 13 out of 23 male mice that spontaneously died or reached pre-defined humane endpoint could have been euthanized on the 3rd day post-MCAo (average time of death = 4.08 (1.07) days). Six out of 181 survived mice (3.3%) were falsely predicted to die. Four out of 10 female mice that spontaneously died or reached pre-defined humane endpoint could have been euthanized on the 2nd day post-MCAo (average time of death = 4.25 (2.28) days). Three out of 77 survived mice (3.9%) were falsely predicted to die (Figure 3.13, a, b). Death could not be predicted with sufficient accuracy until the 2nd (female mice) or 3rd (male mice) day post-MCAo (Table 3.1, a, b).

In the sepsis model, 25 out of 28 animals could have been euthanized 24 hours post-treatment (average time of death = 58.7 (35.0) hours). Six out of 254 survived mice (2.3%) were falsely predicted to die (Figure 3.13, c). Data obtained at 12 hours after the 1st injection could not be used to predict impending death reliably due to the low sensitivity, precision and accuracy (Table 3.1, c).

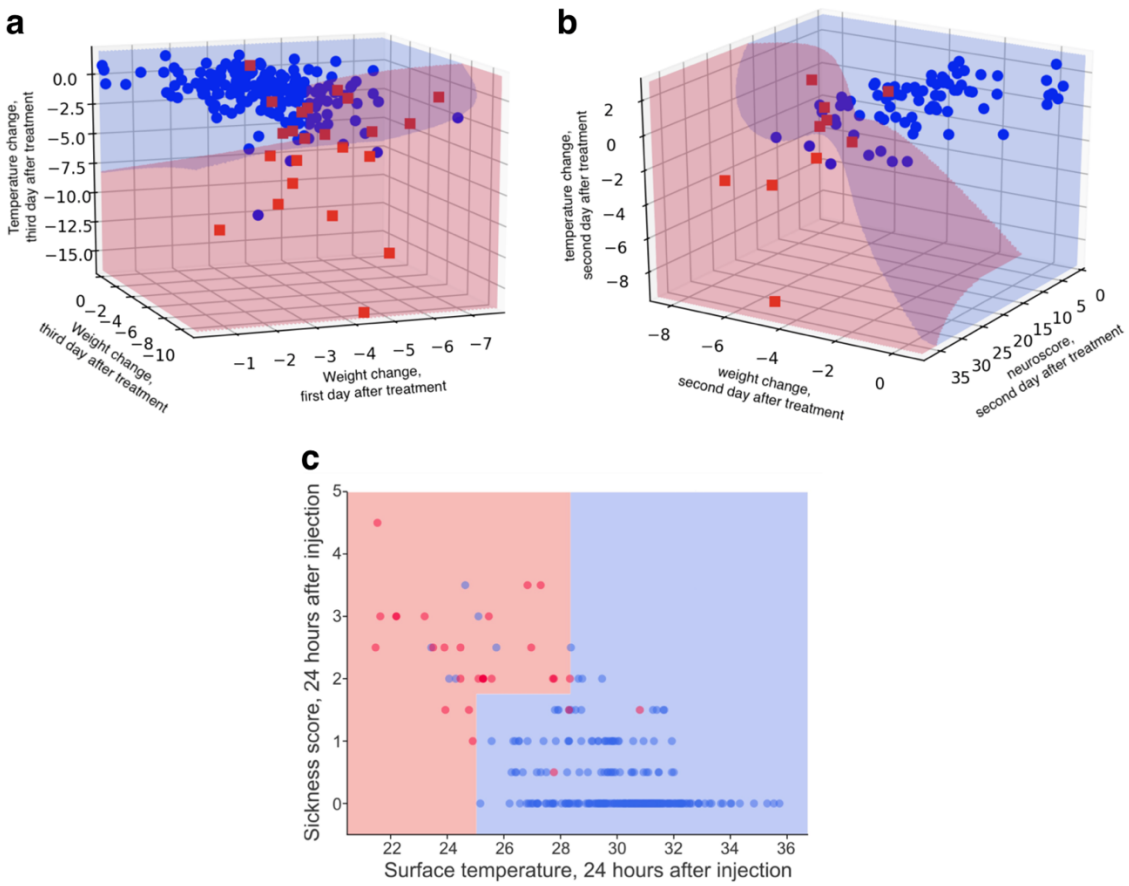


**Figure 3.13:** Using decision boundaries determined by machine learning models in death prediction. (a). Male mice, stroke model. (b). Female mice, stroke model. (c). Sepsis model. Blue dot, survived animal; red dot, animal euthanized upon reaching the pre-defined humane endpoint or died spontaneously; blue area, predicted survival; red area, predicted death.

### 3.2.3. Using additional parameters in model training to improve performance

In male mice of the stroke model, using weight change on the 1st day post-MCAo and core temperature change on the 3rd day post-MCAo as additional parameters increased sensitivity from 0.34 to 0.61, precision from 0.64 to 0.74, and accuracy from 0.91 to 0.93 (Table 3.1, a). In female mice, using neuroscore and core temperature change on the 2nd day post-MCAo in addition to weight change have improved sensitivity from 0.19 to 0.69, precision from 0.25 to 0.83 and accuracy from 0.86 to 0.93 (Table 3.1, b).

In the sepsis model, when the model was trained with data from 24 hours post-treatment, sickness score as an additional parameter in model training increased sensitivity and accuracy by 0.22 and 0.01 (Table 3.1, c). No improvement in model performance was observed when sickness score was used as the additional parameter in model training for data from 36 hours post-treatment.

**a.**

| data | male, stroke model | | |
|---|---|---|---|
| physiological parameters used in model training | weight change, 3rd day after treatment | 1. weight change, 1st day after treatment<br>2. weight change, 2nd day after treatment<br>3. core temperature change, 2nd day after treatment | 1. weight change, 1st day after treatment<br>2. weight change, 3rd day after treatment<br>3. core temperature change, 3rd day after treatment |
| sensitivity | 0.339 (0.148) | 0.482 (0.081) | 0.613 (0.088) |
| precision | 0.644 (0.274) | 0.667 (0.272) | 0.738 (0.070) |
| accuracy | 0.907 (0.026) | 0.902 (0.046) | 0.932 (0.018) |
| averaged | 0.630 | 0.684 | 0.761 |

**b.**

| data | female, stroke model | | |
|---|---|---|---|
| physiological parameters used in model training | weight change, 2nd day after treatment | 1. neuroscore, 1st day after treatment<br>2. weight change, 1st day after treatment<br>3. core temperature change, 1st day after treatment | 1. neuroscore, 2nd day after treatment<br>2. weight change, 2nd day after treatment<br>3. core temperature change, 2nd day after treatment |
| sensitivity | 0.194 (0.142) | 0.361 (0.307) | 0.694 (0.275) |
| precision | 0.25 (0.204) | 0.417 (0.312) | 0.833 (0.236) |
| accuracy | 0.863 (0.045) | 0.896 (0.029) | 0.930 (0.030) |
| averaged | 0.436 | 0.558 | 0.819 |

**c.**

| data | sepsis model, 12 hours after 1st treatment | | sepsis model, 24 hours after 1st treatment | | sepsis model, 36 hours after 1st treatment | |
|---|---|---|---|---|---|---|
| physiological parameters used in model training | surface temperature | surface temperature and sickness score | surface temperature | surface temperature and sickness score | surface temperature | surface temperature and sickness score |
| sensitivity | 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 | 0.303 | 0.303 | 0.789 | 0.857 | 0.818 | 0.818 |

**Table 3.1:** Death prediction with individual or combination of parameters at different time points. (a). Male mice, stroke model. (b). Female mice, stroke model. (c). Sepsis model. Data are shown as mean (SD).

## 3.3. Sub-project 3: Automatization of experiments with 8-arm radial maze

### 3.3.1. Entry to the RAM

All animals entered the automated RAM during the experiment. The average latency to first RAM entry was 17h55 min ± 9h14 min. Seven out of 10 animals entered the automated RAM within the first 1h34 min after the start of the experiment. The average number of RAM visits was 8.81 ± 0.27 times per day during the working memory and 12.40 ± 0.49 times per day during the combined working/reference memory paradigms, respectively. The average

number of RAM visits per hour showed good agreement with the animals' diurnal activity pattern (Figure 3.14).

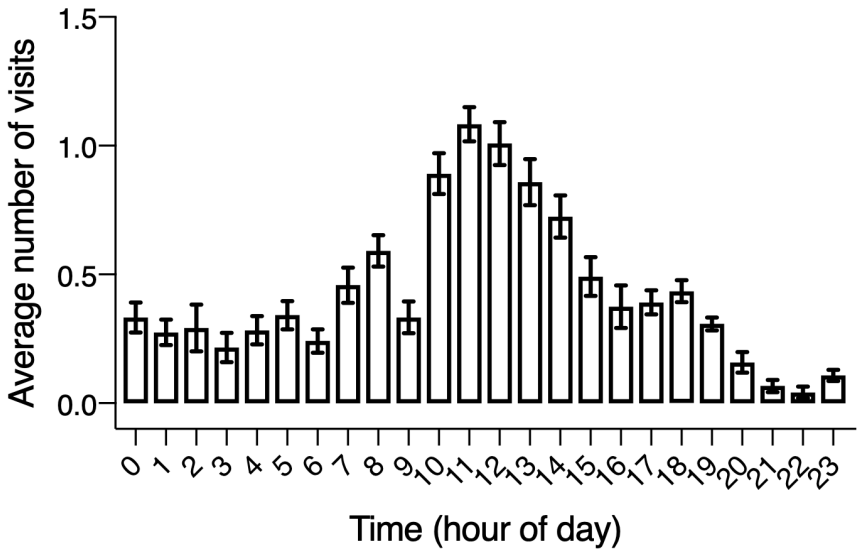


**Figure 3.14:** Average number of visits to the automated radial 8-arm maze by hour across all animals and experiments. Data are presented as mean ± SEM.

### 3.3.2. Working memory paradigm

The number of daily maze visits remained unchanged ($F = 2.65$, $p = 0.11$; Figure 3.15, a), while both the number of arm entries per session and the time needed to complete one session decreased ($F = 6.89$, $p < 0.01$ and $F = 14.62$, $p < 0.001$; Figure 3.15, b, c). The increase in the number of baited arm entries was not significant ($F = 4.77$, $p = 0.05$; Figure 3.15, d) while the number of working memory errors decreased ($F = 7.85$; $p < 0.001$; Figure 3.15, e). The ratio of the number of baited arm entries to total arm entries increased ($F = 10.11$, $p < 0.0001$; Fig. 3.15, f).

### 3.3.3. Combined working/reference memory paradigm

The number of daily visits increased and the number of arm entries per session decreased ($F = 16.13$, $p < 0.0001$ and $F = 4.92$, $p = 0.03$; Figure 3.16, a, b), while the session duration remained unchanged ($F = 0.76$, $p = 0.49$; Figure 3.16, c). No increase was observed in the number of baited arms entered ($F = 1.09$, $p = 0.33$; Figure 3.16, d). There was no effect of time on working memory error ($F = 2.02$, $p = 0.16$; Figure 3.16, e). However, reference memory error and re-entries into un-baited arms decreased ($F = 11.13$, $p < 0.001$ and $F =$

4.06, p < 0.05; Figure 3.16, f, g). The number of baited arms entered in relation to total arm entries remained unchanged (F = 1.85, p = 0.19; Figure 3.16, h).

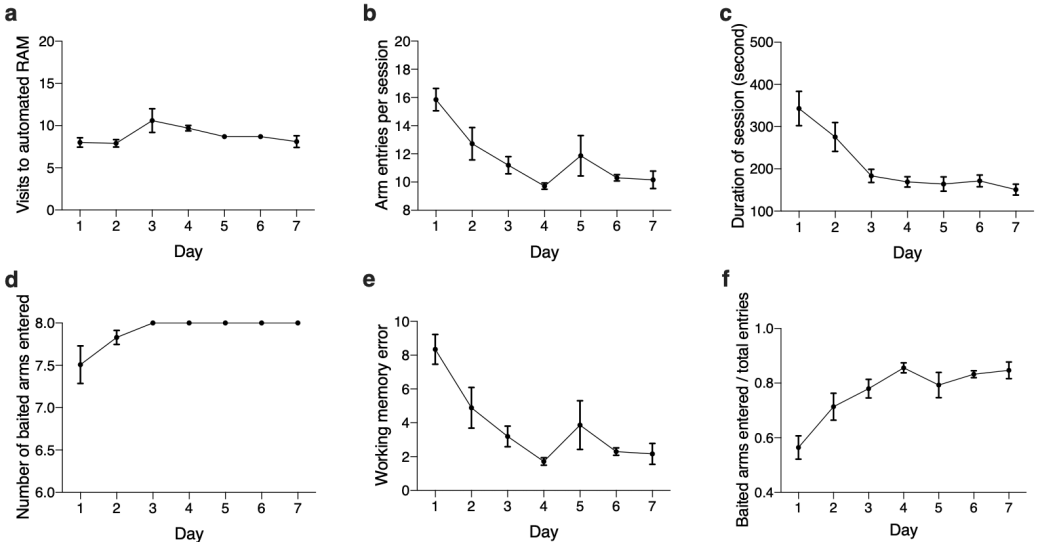


**Figure 3.15:** Learning performance of mice tested with the automated RAM using a spatial working memory paradigm. Data are presented as mean ± SEM.

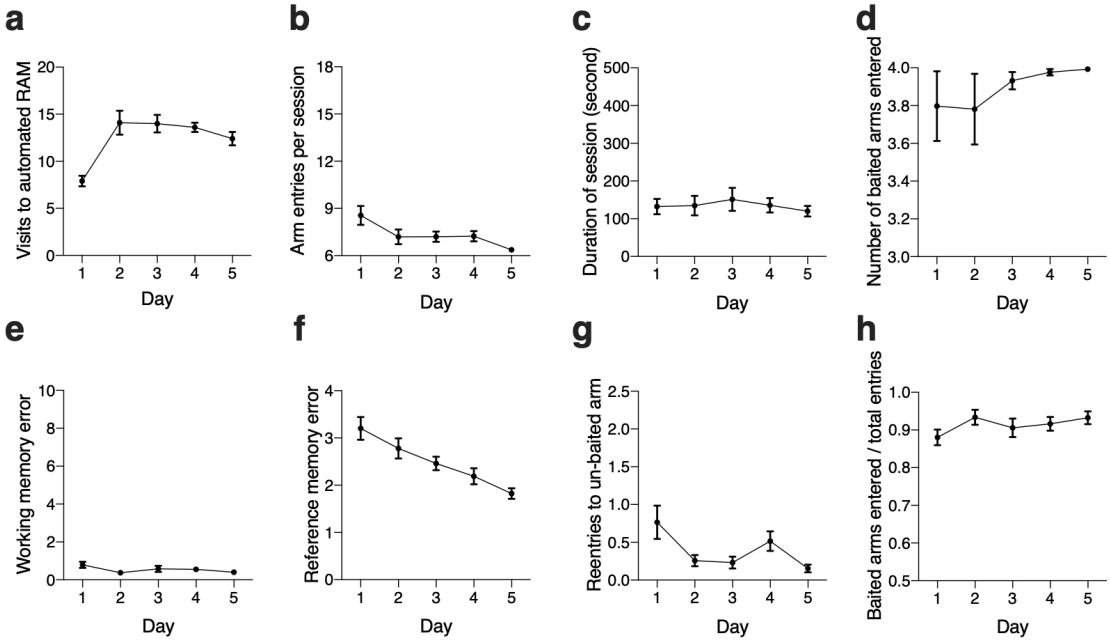


**Figure 3.16:** Learning performance of mice tested with the automated RAM using a combined spatial working/reference memory paradigm. Data are presented as mean ± SEM.

### 3.3.4. Spatial learning by activity phase

The number of daily visits to the automated RAM was reduced during the light phase when compared to the dark phase (3.36 ± 0.24 vs 5.97 ± 0.24 and 3.04 ± 0.23 vs 9.74 ± 0.34 for working memory and combined working/reference memory paradigms, respectively; Figure

3.17, a and Figure 3.18, a). Due to the low number of visits, between and within-subject variation increased during the light phase when compared to the dark phase for working memory and combined working/reference memory paradigms, respectively (Fig. 3.17 and Fig. 3.18). However, despite these variations, there was good general agreement of learning performance between the two phases (Fig. 3.17 and Fig. 3.18).

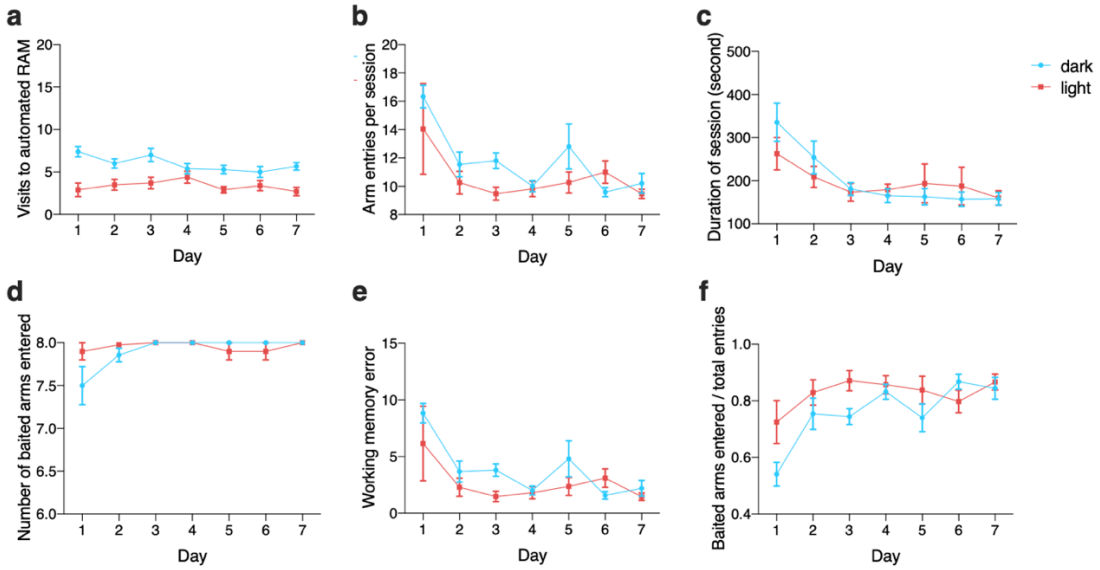


**Figure 3.17:** Learning performance of mice during the working memory paradigm by activity phase. Blue line, dark phase; red line, light phase. Data are presented as mean ± SEM.

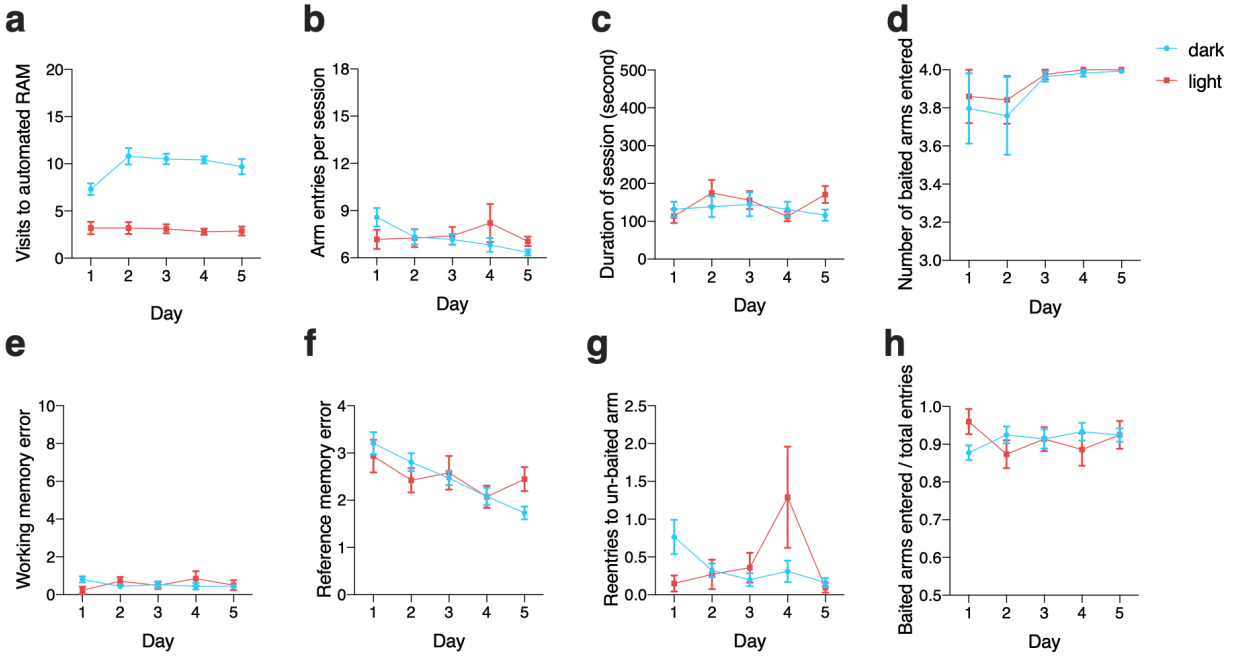


**Figure 3.18:** Learning performance of mice during the combined working/reference memory paradigm by activity phase. Blue line, dark phase; red line, light phase. Data are presented as mean ± SEM.

# 4. Discussion

With the aim to understand how phagocytic deficiency or inhibition of phagocytic signaling pathways may affect cognitive functions, we have (a). Elicited sickness behavior in a mouse model of LPS-induced sepsis, (b). assessed the cognitive functions of mice deficient for Mertk, Cd11b or Mfge8, and (c). assessed the cognitive functions of mice treated with Cilengitide or cRGD. In addition, to maximize reproducibility of research and to minimize unnecessary suffering of animals, attempts have been made to automate an 8-arm radial arm maze for behavioral testing and to systematically determine humane endpoints for detecting animals at higher risk of death at earlier time points.

Using a mouse model of sepsis, the current study showed improved performance in certain behavioral tasks in phagocytosis-deficient animals. In addition, we conducted experiments with the automated 8-arm RAM which required no food/water restriction or manual handling and observed good spatial working and reference memory in mice up to 18 months of age. Lastly, machine learning-based endpoint determination has allowed for animals at higher risk of death to be identified at earlier time points.

## 4.1. Lack of phagocytic proteins or treatment with Cilengitide or cRGD alleviated certain cognitive deficits

In the present study, after a one-month recovery period, reduced activity was no longer present in LPS-inject mice. In alignment with previous research (Arai *et al*., 2001), LPS-treated mice showed compromised performance in place learning and reversal, indicating impaired spatial learning. Another sign of cognitive deficits in LPS-treated mice was the higher preference towards the previously air-puffing corner in avoidance conditioning. Since LPS can decrease sucrose preference and induce long-term anhedonia-like behavior (Painsipp *et al*., 2011; Walker *et al*., 2018), it was unlikely that carry-over effects from the sucrose preference paradigm have caused more frequent nosepokes at the previously air-puffing corner in LPS-injected animals. Therefore, a more plausible explanation of this phenomenon is LPS may impair avoidance retention. In summary, despite recovery in the level of activity from immediate sickness behavior, LPS-injected animals showed signs of long-term cognitive deficits in spatial learning and avoidance retention.

To our knowledge, this is the first study to assess long-term cognitive deficits in phagocytosis-deficient animals. Previous research has linked Mertk- or Mfge8-deficiency with reduced brain atrophy, improved neurological function and alleviated motor deficits (Neher *et al*., 2013). Thus, improved initial spatial learning in mice with Mertk-deficiency could potentially be explained by the prevention of neuronal/synaptic loss and therefore, preserved cognitive functions by Mertk-deficiency. In the meantime, Cilengitide-treated animals displayed improved spatial learning at the beginning of the place learning paradigm compared with animals treated with the inactive control peptide cRGD, which could potentially be attributed to inhibition of phagocytic signaling pathways.

As Neher *et al*., (2013) have used a rodent model of focal brain ischemia and examined the motor functions of animals used in the study, it is unclear whether the same functional implications of Mertk- or Mfge8-deficiency would generalize to a mouse model of sepsis and non-motor tasks. This might be the reason why Mfge8 wildtype mice displayed efficient spatial learning despite poor avoidance retention. Although not within the scope of the PhD project, histological analysis is being performed to understand how neuronal/synaptic loss is impacted by specific phagocytic pathways. With this analysis, a deeper understanding of (a). the occurrence of neuronal/synaptic loss in brain regions including the hippocampus and frontal cortex following LPS and (b). whether the lack of phagocytic proteins modulates neuronal/synaptic loss in these regions and preserves neurons/synapses, could be achieved, giving rise to more comprehensive explanations of the behavioral data.

### 4.2. Use of machine learning models in determination of humane endpoints

Machine learning, a technique used to identify underlying patterns from given datasets that could be generalized to new data, has been applied in studies using animals (Kabra *et al.,* 2013; Han *et al.,* 2018). With the aim to reduce animals' stress and suffering and to maximize reproducibility of experimental results, we have examined the applicability of machine learning in determining humane endpoints.

While traditional approaches in humane endpoint definition suffer from high degrees of heterogeneity which confounds identification of methods in endpoint determination that can be applied to more than one study (Mei *et al*., 2019), machine learning has enabled

identification of case-specific cut-off values across animal models without fundamental change in methodology. With models trained using body weight, sickness severity scores and surface or core temperature data, a significant percentage of animals at higher risk of death could have been euthanized at least one day earlier. To our knowledge, this is the first study using machine learning models 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 machine learning-based approach for the endpoint determination include improved standardization and comparability, as standardized metrics could be used across studies to evaluate model performance. Moreover, once a model is trained with sufficient data, it could be used repeatedly by researchers and animal technicians to determine whether an animal has an increased risk of death.

### 4.3. Use of automated 8-arm radial maze in the study of spatial learning

Attempts have been made to (a). automatize the recording of experiment-related events by using pellet sensors (Miyakawa *et al*., 2001), photo emitter/detector pairs (Peele and Baron, 1988), and pressure detectors (Dubreuil *et al*., 2003) allowing for remote monitoring of an animals' location and learning performance, (b). control mechanical parts of the RAM without requiring direct manual manipulation by using automated doors (Brillaud, Morillion and De Seze, 2005), and (c). facilitate data collection and visualization by automatically recording an animal's reward intake (Miyakawa *et al*., 2001). However, to the best of our knowledge, none of the previous approaches allowed for the absence of a human experimenter and voluntary 24/7 entry of animals into the apparatus without requiring any food and/or water deprivation. In this study, with the custom-made habituation apparatus and the automated 8-arm RAM, animals have voluntarily entered the RAM and achieved efficient spatial learning without food and/or water restriction. Automatization of the RAM was achieved on three levels: (a). handling-free habituation/experimental procedures, (b). online data collection and (c). offline data analysis.

Mice could achieve efficient spatial learning within the first 3-4 days of both paradigms. In working memory paradigm, working memory error has decreased while the time for session completion has shortened. In combined working/reference memory paradigm, animals showed significant decrease in reference memory error and re-entries into un-baited

arms. In the working memory paradigm of Miyakawa *et al.,* (2001) using a partially automated RAM, mice generally required more than 10 daily trials to achieve a revisiting error less than 10, and the time that animals spent to collect 8 pellets was the shortest on the 13$^{th}$ day (~400 seconds). In the present study, the time required to complete one session was 204.56 seconds on day 7 with an average of 7.88 pellets collected per session. A possible explanation of the improved spatial learning performance of mice in the automated 24/7 accessible RAM is the markedly increased daily visit frequency. Experiments in the automated RAM were carried out without the presence of experimenters, therefore, the automated RAM may improve standardization and comparability of results across studies by excluding potential experimenter effects and bias.

## 4.4. Limitations, outlook and conclusions

### 4.4.1. Sub-project 1: Assessment of physiological and cognitive alterations in the mouse model of sepsis, and the effect of phagocytic deficiency or inhibition of phagocytic signaling pathways on cognitive functions

One of the limiting factors in sickness behavior monitoring was the proximity of the RFID transponder's implantation site to brown adipose tissue located in the interscapular region, which is an important heat generator in mice and thus might have a confounding effect on core temperature readings (Mei *et al.*, 2018). Secondly, due to transponder malfunction and technical errors during temperature acquisition, core temperature measurements from 22 animals were partially or fully excluded from data analysis.

As the large number of parameters in experiments with IntelliCage could be particularly challenging for interpretation and replication of experimental results, we have tried to avoid introducing novel parameters and primarily examined those frequently used in published studies. In the present experimental design, experimenter presence and manual handling were unavoidable during sucrose preference and avoidance conditioning, when sucrose solution was introduced, or animals were transported from and back to the homecage. This intervention, in combination with the presentation of aversive stimulus during avoidance conditioning, can introduce a source of bias, as mice have been housed in the IntelliCage for 2 weeks without direct contact with the experimenter. Secondly, the use of sucrose solution prior to paradigms involving aversive stimulus may confound the

experimental results. Given these limitations, it would be desirable to re-arrange the experimental paradigms to optimize the protocol. For example, the sucrose preference paradigm could be conducted shortly after animals have been familiarized with IntelliCage, instead of by the end of the experiment, to minimize the confounding effects of sucrose-induced hedonic behavior on avoidance conditioning. This way, drinking session adaptation and the spatial learning paradigms could be conducted between the sucrose preference paradigm and avoidance conditioning, introducing a period of 10 days to wash out the effect of sucrose preference. Thirdly, due to technical problems with genotyping, 7 animals (1 Mertk, 6 Cd11b) were excluded from data analysis, which led to a small sample size and potential difficulties in identifying significant relationships. Furthermore, hardware malfunctions of the IntelliCage including delayed registration of nosepokes and licks had been a recurring issue during device testing, and it would be necessary to avoid the same technical problems in future studies. As mice brain tissue from the present study is still being examined, behavioral data will be re-interpreted with care when results from immunohistochemistry and synaptosome analysis are made available. It would be essential, however, to conduct future studies to further quantify the effects of phagocytic deficiency or inhibition of phagocytic signaling pathways on behavior and cognition.

### 4.4.2. Sub-project 2: Refinement of humane endpoints in animal models of acute disease

Although a machine learning model has been tested in the present study, the availability of data that could be used to train and validate the current model was limited, as few studies have made raw data available in open data repositories. Therefore, the current model is of low complexity and may not generalize well in studies of larger populations. A future step would be to make the model public and obtain data from laboratories that use the model, to train the model again based on datasets of larger-scale, and to test the generalizability of the present method in other disease models.

### 4.4.3. Sub-project 3: Automatization of experiments with 8-arm radial maze

Before being applied in future studies, limitations of the current automated RAM setup should be assessed carefully. In the present study, 4.19% of data entries were excluded due to

hardware or software malfunctions, indicating the need for further optimization of the automated RAM. Secondly, it was not possible to determine the number of pellets being eaten per session with the current setup. This issue could be addressed by adding sensors or video-based analysis tools to the automated RAM. Thirdly, 4 animals were used to examine the functionality of the RAM at 9 months of age and then tested again at 18 months, which might be a source of bias. In addition, the number of daily entries to the automated RAM was not limited and may have resulted in some animals occupying the automated RAM, potentially outperforming their littermates. Although no animal significantly outperformed their littermates in the present study, in future studies, a pre-defined cut-off value for the daily visit frequency could be applied to ensure all animals receive comparable training in the automated RAM.

### 4.4.4. Conclusion

In the present study, phagocytic deficiency for Mertk, Cd11b or Mfge8 or treatments with peptides have led to improved cognitive performance in certain behavioral tasks. Improved performance in behavioral tasks by the lack of phagocytic proteins or inhibition of phagocytic signaling are strain- and treatment-specific. Machine learning-based humane endpoint determination, together with automated behavioral testing systems, may minimize suffering and discomfort for experimental animals and optimize the interpretability and reproducibility of experimental results. Taken together, results from the current study indicated that (a). inhibition of specific phagocytic pathways may be a potential therapeutic target to alleviate cognitive deficits following critical illness, (b). machine learning-based endpoint determination and a fully automated cognitive testing apparatus minimized unnecessary suffering and handling-induced stress in animal experiments. Nevertheless, cognitive data should be interpreted in combination with histological analysis, and it would be essential to conduct more studies to further quantify the effect of phagocytic deficiency on cognitive functions. Limitations of both machine learning-based endpoint determination and the fully automated 8-arm RAM should be carefully addressed before being applied in animal experimentation.

# 5. References


1. Anderson, S. T., Commins, S., Moynagh, P. N. & Coogan, A. N. 2015. Lipopolysaccharide-induced sepsis induces long-lasting affective changes in the mouse. Brain, behavior, and immunity, 43, 98-109.
2. Arai, K., Matsuki, N., Ikegaya, Y. and Nishiyama, N., 2001. Deterioration of spatial learning performances in lipopolysaccharide-treated mice. The Japanese journal of pharmacology, 87(3), pp.195-201.
3. Brillaud, E., Morillion, D. and De Seze, R., 2005. Modest environmental enrichment: effect on a radial maze validation and well being of rats. Brain research, 1054(2), pp.174-182.
4. Brown, G. C. & Neher, J. J. 2012. Eaten alive! Cell death by primary phagocytosis: 'phagoptosis'. Trends in biochemical sciences, 37(8), pp 325-332.
5. Brown, G. C. & Neher, J. J. 2014. Microglial phagocytosis of live neurons. Nature Reviews Neuroscience, 15(4), pp 209.
6. Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W. & Kelley, K. W. 2008. From inflammation to sickness and depression: when the immune system subjugates the brain. Nature reviews neuroscience, 9(1), pp 46.
7. Deacon, R. M. 2006. Assessing nest building in mice. Nature protocols, 1(3), pp 1117.
8. 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. Nature Precedings, doi:10.1038/npre.2012.3492.3
9. Donath, S., An, J., Lee, S.L.L., Gertz, K., Datwyler, A.L., Harms, U., Müller, S., Farr, T.D., Füchtemeier, M., Lättig-Tünnemann, G. and Lips, J., 2016. Interaction of ARC and Daxx: A novel endogenous target to preserve motor function and cell loss after focal brain ischemia in mice. Journal of Neuroscience, 36(31), pp.8132-8148.
10. Dubreuil, D., Tixier, C., Dutrieux, G. and Edeline, J.M., 2003. Does the radial arm maze necessarily test spatial memory?. Neurobiology of learning and memory, 79(1), pp.109-117.
11. Dzik, J. M., Puścian, A., Mijakowska, Z., Radwanska, K. & Łęski, S. 2018. PyMICE: APython library for analysis of IntelliCage data. Behavior research methods, 1-12.

12. Emmrich, J.V., Neher, J.J., Boehm-Sturm, P., Endres, M., Dirnagl, U. and Harms, C., 2017. Stage 1 Registered Report: Effect of deficient phagocytosis on neuronal survival and neurological outcome after temporary middle cerebral artery occlusion (tMCAo). F1000Research, 6.
13. Fricker, M., Neher, J. J., Zhao, J.-W., Théry, C., Tolkovsky, A. M. & Brown, G. C. 2012. MFG-E8 mediates primary phagocytosis of viable neurons during neuroinflammation. Journal of Neuroscience, 32(8), pp 2657-2666.
14. Han, S., Taralova, E., Dupre, C. & Yuste, R. 2018. Comprehensive machine learning analysis of Hydra behavior reveals a stable basal behavioral repertoire. elife, 7, e32605.
15. Hurst, J.L. and West, R.S., 2010. Taming anxiety in laboratory mice. Nature methods, 7(10), p.825.
16. Johansson, A., Fredriksson, R., Winnergren, S., Hulting, A.-L., Schiöth, H. B. & Lindblom, J. 2008. The relative impact of chronic food restriction and acute food deprivation on plasma hormone levels and hypothalamic neuropeptide expression. Peptides, 29(9), pp 1588-1595.
17. Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. 2013. JAABA: interactive machine learning for automatic annotation of animal behavior. Nature methods, 10(1), pp 64.
18. Lemstra, A. W., Hoozemans, J. J., van Haastert, E. S., Rozemuller, A. J., Eikelenboom, P. & van Gool, W. A. 2007. Microglia activation in sepsis: a case-control study. Journal of neuroinflammation, 4(1), pp 4.
19. Linnartz, B., Kopatz, J., Tenner, A. J. & Neumann, H. 2012. Sialic acid on the neuronal glycocalyx prevents complement C1 binding and complement receptor-3-mediated removal by microglia. Journal of Neuroscience, 32(3), pp 946-952.
20. Lipp, H.P., Litvin, O., Galsworthy, M., Vyssotski, D.L., Zinn, P., Rau, A.E., Neuhausser-Wespy, F., Wurbel, H., Nitsch, R. and Wolfer, D.P., 2005. Automated behavioral analysis of mice using INTELLICAGE: inter-laboratory comparisons and validation with exploratory behavior and spatial learning. In Proceedings of Measuring Behavior 2005 (pp. 66-69). Noldus Information Technology: Netherlands.

21. Mei, J., Riedel, N., Grittner, U., Endres, M., Banneke, S. & Emmrich, J. V. 2018. Body temperature measurement in mice during acute illness: implantable temperature transponder versus surface infrared thermometry. Scientific reports, 8(1), pp 3526.
22. Mei, J., Banneke, S., Lips, J., Kuffner, M., Hoffmann, C., Dirnagl, U., Endres, M., Harms, C. and Emmrich, J. 2019. Refining humane endpoints in mouse models of disease by systematic review and machine learning-based endpoint definition. ALTEX - Alternatives to animal experimentation.
23. Miyakawa, T., Yamada, M., Duttaroy, A. & Wess, J. 2001. Hyperactivity and intact hippocampus-dependent learning in mice lacking the M1 muscarinic acetylcholine receptor. Journal of Neuroscience, 21(14), pp 5239-5250.
24. Neher, J. J., Emmrich, J. V., Fricker, M., Mander, P. K., Théry, C. & Brown, G. C. 2013. Phagocytosis executes delayed neuronal death after focal brain ischemia. Proceedings of the National Academy of Sciences, 110(43), pp E4098-E4107.
25. Olton, D. S. & Samuelson, R. J. 1976. Remembrance of places passed: spatial memory in rats. Journal of Experimental Psychology: Animal Behavior Processes, 2(2), pp 97.
26. Painsipp, E., Köfer, M.J., Sinner, F. and Holzer, P., 2011. Prolonged depression-like behavior caused by immune challenge: influence of mouse strain and social environment. PLoS One, 6(6), p.e20719.
27. Pandharipande, P. P., Girard, T. D., Jackson, J. C., Morandi, A., Thompson, J. L., Pun, B. T., Brummel, N. E., Hughes, C. G., Vasilevskis, E. E. & Shintani, A. K. 2013. Long-term cognitive impairment after critical illness. New England Journal of Medicine, 369(14), pp 1306-1316.
28. Peele, D.B. and Baron, S.P., 1988. Effects of selection delays on radial maze performance: Acquisition and effects of scopolamine. Pharmacology Biochemistry and Behavior, 29(1), pp.143-150.
29. Pesic, V., Marinkovic, P., Janac, B., Ignjatovic, S., Popic, J., Kanazir, S. & Ruzdijic, S. 2010. Changes of behavioral parameters during long-term food restriction in middle-aged Wistar rats. Physiology & behavior, 101(5), pp 672-678.
30. Ravichandran, K. S. 2011. Beginnings of a good apoptotic meal: the find-me and eat-me signaling pathways. Immunity, 35(4), pp 445-455.

31. Reberg, D., Mann, B. & Innis, N. K. 1977. Superstitious behavior for food and water in the rat. Physiology & behavior, 19(6), pp 803-806.
32. Russell, W. M. S., Burch, R. L. & Hume, C. W. 1959. The principles of humane experimental technique. Methuen London.
33. Schafer, D. P., Lehrman, E. K., Kautzman, A. G., Koyama, R., Mardinly, A. R., Yamasaki, R., Ransohoff, R. M., Greenberg, M. E., Barres, B. A. & Stevens, B. 2012. Microglia sculpt postnatal neural circuits in an activity and complement-dependent manner. Neuron, 74(4), pp 691-705.
34. Schellinck, H.M., Cyr, D.P. and Brown, R.E., 2010. How many ways can mouse behavioral experiments go wrong? Confounding variables in mouse models of neurodegenerative diseases and how to control them. In Advances in the Study of Behavior (Vol. 41, pp. 255-366). Academic Press.
35. Semmler, A., Hermann, S., Mormann, F., Weberpals, M., Paxian, S. A., Okulla, T., Schäfers, M., Kummer, M. P., Klockgether, T. & Heneka, M. T. 2008. Sepsis causes neuroinflammation and concomitant decrease of cerebral metabolism. Journal of neuroinflammation, 5(1), pp 38.
36. Semmler, A., Widmann, C. N., Okulla, T., Urbach, H., Kaiser, M., Widman, G., Mormann, F., Weide, J., Fliessbach, K. & Hoeft, A. 2013. Persistent cognitive impairment, hippocampal atrophy and EEG changes in sepsis survivors. J Neurol Neurosurg Psychiatry, 84(1), pp 62-69.
37. Vorhees, C. V. & Williams, M. T. 2014. Assessing spatial learning and memory in rodents. Ilar Journal, 55(2), pp 310-332.
38. Walker, A.K., Wing, E.E., Banks, W.A. and Dantzer, R., 2018. Leucine competes with kynurenine for blood-to-brain transport and prevents lipopolysaccharide-induced depression-like behavior in mice. Molecular psychiatry, p.1.
39. Widmann, C. N. & Heneka, M. T. 2014. Long-term cerebral consequences of sepsis. The Lancet Neurology, 13(6), pp 630-636.

## Eidesstattliche Versicherung

„Ich, Jie Mei, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Sepsis-associated cognitive dysfunction: An investigation using stress-free, automated behavioral tests“ (German translation: Sepsis-assoziierte kognitive Dysfunktion: Eine Untersuchung mit stressfreien, automatisierten Verhaltenstests) selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

_____

Datum Unterschrift

# Anteilserklärung an der erfolgten Publikation

Jie Mei has contributed to the following publications during her doctoral candidacy at Department of Neurology and Department of Experimental Neurology, Charité – Universitätsmedizin Berlin:

**Publikation**: Mei, J., Banneke, S., Lips, J., Kuffner, M., Hoffmann, C., Dirnagl, U., Endres, M., Harms, C. and Emmrich, J. Refining humane endpoints in mouse models of disease by systematic review and machine learning-based endpoint definition. *ALTEX - Alternatives to animal experimentation. doi: 10.14573/altex.1812231*. 2019.

Beitrag im Einzelnen:
Jie Mei has

(a) collected, cleaned and organized all data for the animal model of sepsis:
    a. Data collection, pre-processing and cleaning before predictive analysis was performed by Jie Mei (Tab. 1), including physiological data of body weight, body temperature, and sickness severity score; Sickness behavior monitoring with physiological data acquisition and collection (including body weight, body temperature, and sickness severity score) was performed by Jie Mei, with the help of animal technicians;

(b) cleaned and organized all data for the animal model of stroke:
    b. Data pre-processing and cleaning before predictive analysis was performed by Jie Mei (Tab. 1), including physiological data of body weight, body temperature, and modified DeSimoni neuroscore;

(c) designed and implemented data analysis methods:
    c. Scripts for machine learning models, data import and extraction and plotting were written by Jie Mei (Tab. 4; Fig. 2); Manual data inspection and pre-processing were conducted by Jie Mei;

(d) conducted data analysis:
    a. Statistical analysis in SPSS and Python was performed by Jie Mei, including descriptive statistics (Tab. 3), machine-learning based prediction of impending death (Tab. 4) and generating of decision boundaries (Fig. 2);

(e) conducted literature search and title, abstract and full text screening:
    a. All reviewed papers were searched, screened and selected by Jie Mei under the supervision of Dr. Julius Emmrich (Fig. 1; Tab. 2);

(f) produced all tables (Tab 1-5), figures (Fig. 1 and Fig. 2) and contributed to writing the manuscript:
    a. Jie Mei has drafted and revised the manuscript, generated all figures and produced all tables under the supervision of Dr. Julius Emmrich and Prof. Christoph Harms.

_____

Unterschrift, Datum und Stempel der betreuenden Hochschullehrer/der betreuenden Hochschullehrerinnen

_____

Unterschrift des Doktoranden/der Doktorandin

Journal Data Filtered By: **Selected JCR Year: 2017** Selected Editions: SCIE,SSCI
Selected Categories: **"MEDICINE, RESEARCH and EXPERIMENTAL"**
Selected Category Scheme: WoS
**Gesamtanzahl: 133 Journale**

| Rank | Full Journal Title | Total Cites | Journal Impact Factor | Eigenfactor Score |
|---|---|---|---|---|
| 1 | NATURE MEDICINE | 75,461 | 32.621 | 0.171980 |
| 2 | Science Translational Medicine | 26,691 | 16.710 | 0.126450 |
| 3 | Annual Review of Medicine | 6,111 | 14.970 | 0.010320 |
| 4 | JOURNAL OF CLINICAL INVESTIGATION | 107,818 | 13.251 | 0.165270 |
| 5 | TRENDS IN MOLECULAR MEDICINE | 9,213 | 11.021 | 0.019720 |
| 6 | JOURNAL OF EXPERIMENTAL MEDICINE | 62,537 | 10.790 | 0.078310 |
| 7 | EMBO Molecular Medicine | 6,402 | 10.293 | 0.026160 |
| 8 | Theranostics | 5,761 | 8.537 | 0.015710 |
| 9 | MOLECULAR ASPECTS OF MEDICINE | 5,157 | 7.344 | 0.009700 |
| 10 | MOLECULAR THERAPY | 16,013 | 7.008 | 0.029180 |
| 11 | Nanomedicine-Nanotechnology Biology and Medicine | 9,338 | 6.500 | 0.015670 |
| 12 | Wiley Interdisciplinary Reviews-Nanomedicine and Nanobiotechnology | 2,088 | 6.350 | 0.004260 |
| 13 | EBioMedicine | 3,378 | 6.183 | 0.015290 |
| 14 | Molecular Therapy-Nucleic Acids | 2,180 | 5.660 | 0.008200 |
| 15 | EXPERIMENTAL AND MOLECULAR MEDICINE | 3,538 | 5.584 | 0.007100 |
| 16 | ALTEX-Alternatives to Animal Experimentation | 1,261 | 5.232 | 0.001990 |
| 17 | CLINICAL SCIENCE | 10,321 | 5.220 | 0.013630 |
| 18 | mAbs | 3,767 | 5.165 | 0.011220 |
| 19 | Stem Cell Research & Therapy | 4,578 | 4.963 | 0.012630 |
| 20 | JOURNAL OF MOLECULAR MEDICINE-JMM | 7,120 | 4.938 | 0.011400 |
| 21 | Translational Research | 3,416 | 4.880 | 0.009000 |
| 22 | XENOTRANSPLANTATION | 1,479 | 4.717 | 0.002550 |
| 23 | Cancer Biology & Medicine | 816 | 4.607 | 0.002330 |
| 24 | MOLECULAR PHARMACEUTICS | 15,754 | 4.556 | 0.029080 |
| 25 | JOURNAL OF CELLULAR AND MOLECULAR MEDICINE | 10,938 | 4.302 | 0.015770 |
| 26 | LABORATORY INVESTIGATION | 10,461 | 4.254 | 0.010460 |
| 27 | HUMAN GENE THERAPY | 5,559 | 4.241 | 0.007690 |

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 1$^{st}$, 2$^{nd}$, 7$^{th}$, 14$^{th}$, and 21$^{st}$ 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 2$^{nd}$ criteria, or a score of 3 or 4 of the 3$^{rd}$-12$^{th}$ 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 30$^{th}$ 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,

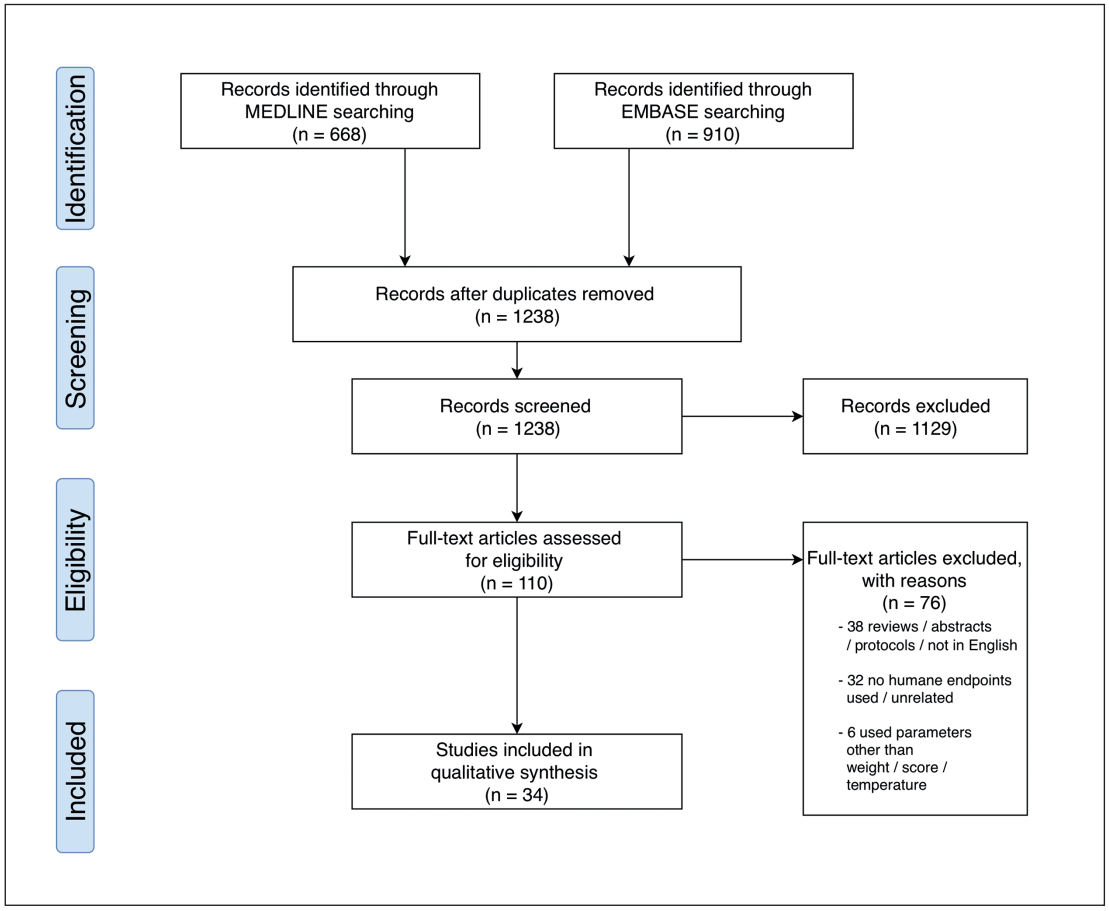


**Fig. 1: Flow diagram showing the number of studies identified, screened, extracted and included in this systematic review**

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; T x BW < 90% of baseline value on day 2 or day 5 after infection; 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; n(sur) = 22 | ≤ 25 days | core temperature | every 15 minutes | core temperature < 23.4°C | Vlach et al., 2000 |
| Fungal infection | n = 160; 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 hu-

mane 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).

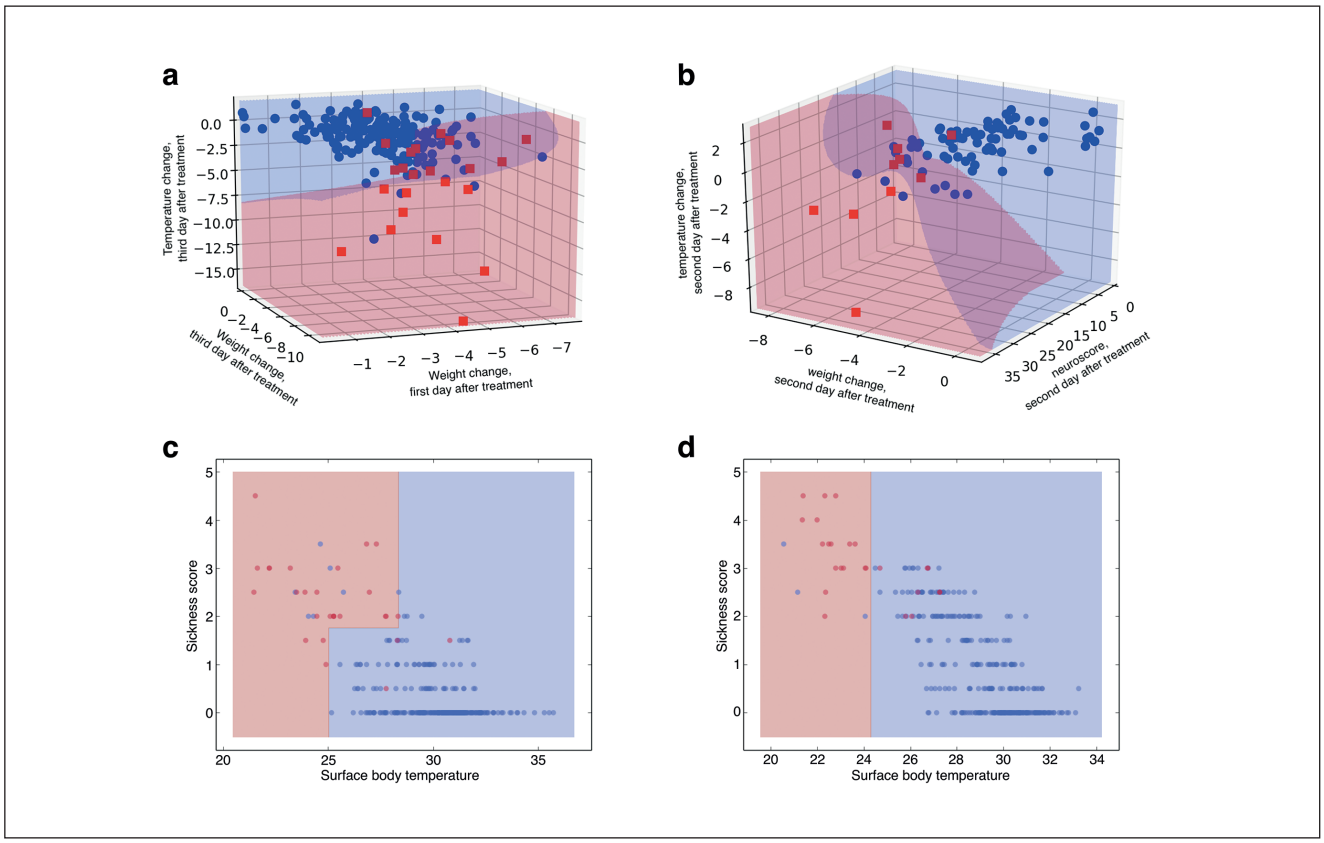


**Fig. 2: Decision boundaries determined by the machine learning model**

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 func-

tion 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.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht

# Publication List

**Publications related to the PhD project:**

- **Mei, J.,** Kohler, J., Spies, C., Endres, M., Banneke, S., and Emmrich, J.V. Automated radial 8-arm maze: A voluntary and stress-free behavior test to assess spatial learning and memory in mice. Under review. (Behavioural Brain Research; BBR_2019_1178).
- Kohler, J., **Mei, J.,** Banneke, S., Endres, M. and Emmrich, J.V. Assessing spatial learning and memory in mice: Automated versus manual radial 8-arm maze. In preparation.
- **Mei, J.**, Banneke, S., Lips, J., Kuffner, M., Hoffmann, C., Dirnagl, U., Endres, M., Harms, C. and Emmrich, J., 2019. Refining humane endpoints in mouse models of disease by systematic review and machine learning-based endpoint definition. ALTEX - Alternatives to animal experimentation. doi: 10.14573/altex.1812231. (Impact Factor: **6.183**)
- **Mei, J.**, Riedel, N., Grittner, U., Endres, M., Banneke, S. and Emmrich, J.V., 2018. Body temperature measurement in mice during acute illness: implantable temperature transponder versus surface infrared thermometry. Scientific reports, 8(1), p.3526. (Impact Factor: **4.011**)

**Other publications:**

- Akgun, O. C., and **Mei, J.**, 2019. An Energy Efficient Time-Mode Digit Classification Neural Network. Philosophical Transactions of the Royal Society A (RSTA-2019-0163). Accepted. (Impact Factor: **3.093**)
- **Mei, J.** and Singh, T., 2018. Intra-thalamic and thalamocortical connectivity: potential implication for deep learning. In Proceedings of the 1st International Workshop on Software Engineering for Cognitive Services (pp. 9-14). ACM.
- Jarmolowska, J., Turconi, M.M., Busan, P., **Mei, J.** and Battaglini, P.P., 2013. A multimenu system based on the P300 component as a time saving procedure for communication with a brain-computer interface. Frontiers in neuroscience, 7, p.39. (Impact Factor: **3.648**)

# Acknowledgements

First of all, I would like to thank my supervisors Prof. Dr. Ulrich Dirnagl, Dr. med. Julius Emmrich and Prof. Dr. Christoph Harms for offering me this great opportunity to work with them and complete my doctoral work. It was their motivation, support and supervision that made the realization of many scientific ideas possible. I thank all my colleagues at Charité – Universitätsmedizin Berlin and Berlin Institute of Health, especially Prof. Dr. Matthias Endres, Dr. Ulrike Grittner, Dr. Nico Riedel, Dr. Vince Madai, Dr. Michelle Livne and Dr. Ahmed Khalil, for their supervision, guidance and support.

I am thankful to Dr. Stefanie Banneke and Prof. Dr. Lars Lewejohann at the Bundesinstitut für Risikobewertung, for providing excellent experimental facilities as well as their supervision. I thank all animal technicians from the Center for the Protection of Laboratory Animals at the Bundesinstitut für Risikobewertung, especially Mr. Paolo Rosellini Tognetti, Ms. Lisa Gordijenko and Mr. David Schreiber, for their technical assistance.

My special thanks go to Mr. Ralf Ansorg and Dr. Torben Hager for their help at various stages.

Finally, I'd like to thank my friends Julien, Shannon, Momo, Lingjie, Ryan and Dr. Arnab Chakrabarty for the wonderful journey.