Ulm University
Institute of Epidemiology and Medical Biometry
Director: Prof. Dr. med. Dietrich Rothenbacher

Designing translational animal experiments by using Bayesian predictive approaches

Dissertation submitted to obtain the doctoral degree of Human Biology
of the Medical Faculty of Ulm University

Theresa Unseld
Ulm
2023

Acting dean: Prof. Dr. Thomas Wirth

First reviewer: Prof. Dr. Benjamin Mayer

Second reviewer: Prof. Dr. Frank Konietschke

Day of graduation: 14th June 2024

Contents

List of Abbreviations v
List of Figures vii
List of Tables ix
List of Examples x
List of Summaries xi
List of Main Results xii
Publications as part of this thesis xiii

1 Introduction **1**

2 Material and Methods **8**
2.1 Classical sample size determination in animal experiments 9
2.1.1 Concepts of frequentist hypothesis testing 10
2.1.2 Problems in the design and analysis of translational animal experiments . 11
2.2 Bayesian predictive approaches 15
2.2.1 Analysis of historical control data 15
2.2.2 Bayesian analysis of animal experiments 21
2.2.3 Choice of prior distributions 22
2.3 Simulation based design analysis and sample size determination 42
2.3.1 Concepts of Bayesian hypothesis testing 42
2.3.2 Estimating power and determining sample size by simulation 52
2.3.3 Additional design quality criteria 57
2.4 Real-world example "bone quality under postmenopausal osteoporosis" 63
2.4.1 Historical data 64
2.4.2 Fitted Bayesian models 67
2.5 Fictive multi-experiment example 73
2.5.1 Bayesian model of the historical and the new data 73
2.5.2 Simulation of prior-data conflicts 76
2.6 Computation in R 76

3 Results **78**
3.1 Results for the outcome bending stiffness 79
3.1.1 Descriptive analysis of the historical data 79
3.1.2 Results from the Bayesian model of the historical control data 79
3.1.3 Design analysis and sample size determination 81
3.2 Results for the outcome relative bone volume 86
3.2.1 Descriptive analysis of the historical data 86
3.2.2 Results from the Bayesian model of the historical control data 86
3.2.3 Design analysis and sample size determination 88
3.3 Total sample size for the real-world example 92
3.4 Results for the fictive multi-experiment example 93
3.4.1 Descriptive analysis of the historical data 93
3.4.2 Results from the Bayesian model of the historical control data 94
3.4.3 Design analysis and sample size determination 95

4 Discussion **100**
4.1 Discussion of the materials and methods 100
4.1.1 Analysis of the historical control data 100
4.1.2 Sample size determination for the new experiment via design simulation . 110
4.2 Discussion of the results 121
4.2.1 Sample size results 121
4.2.2 Results from the investigation of additional design quality criteria 126
4.2.3 Results from the investigated prior-data conflicts 127
4.3 Conclusions 130
4.4 Outlook 133

5 Summary **135**

6 References **137**

Acknowledgments **155**

Curriculum Vitae **156**

List of Abbreviations

ADMA	aggregate data meta-analysis
AIC	Akaike information criterion
ALS	amyotrophic lateral sclerosis
BF	Bayes factor
BV/TV	bone volume to tissue volume
C5aR1	complement component 5a receptor 1
CC	Collaborative Cross
EM	expectation-maximization
ESS	effective sample size
ETI	equal-tailed interval
F1	first filial
F2	second filial
FPR	false positive rate
HC	heteroscedasticity consistent
HDI	highest posterior density interval
IADMA	individual animal data meta-analysis
IMPC	International Mouse Phenotyping Consortium
IPDMA	individual patient data meta-analysis
MA	meta-analysis
MAC	meta-analytic-combined
MAP	meta-analytic-predictive

MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MGI	Mouse Genome Informatics
micro-CT	micro-computed tomography
ML	maximum likelihood
MPD	Mouse Phenome Database
MSE	mean square error
NNHM	normal-normal hierarchical model
3R	Replacement, Reduction and Refinement
RMSE	root mean square error
SD	standard deviation
SE	standard error
SOD1	superoxide dismutase 1
FPR	false positive rate
UIP	unit information prior

List of Figures

1 Historical model part of the sample size planning procedure 16
2 Bayesian estimates of the historical control means (fictive example) 20
3 Conjugate prior illustration (fictive example) 25
4 Posterior distributions (fictive example) 28
5 Prior distributions in the meta-analysis of the historical data (fictive example) . . . 32
6 Prior predictive check (fictive example) 33
7 Prediction part of the sample size planning procedure 34
8 Predicted mean and standard deviation (fictive example) 38
9 Parametric approximation of the predictive distributions (fictive example) 41
10 Hypothesis testing part of the sample size planning procedure 43
11 Credible interval illustration (fictive example) 46
12 Savage-Dickey density ratio (fictive example) 51
13 Design simulation part of the sample size planning procedure 53
14 Sampled data in the first run of the design simulation (fictive example) 55
15 Posterior distributions in the first run of the design simulation (fictive example) . . 56
16 Bayes factor in the first run of the design simulation (fictive example) 56
17 Sample size planning procedure with Bayesian predictive approaches 61
18 Bending stiffness curves 66

19 Parametric approximation of the predictions (bending example) 81
20 Parametric approximation of the predicted distributions (micro-CT example) 87

List of Tables

1 Historical data in a single historical experiment (fictive example) 13
2 Historical data in a second historical experiment (fictive example) 14
3 Assumptions on the relatedness of the new and historical parameters 16
4 Historical multi-experiment control data (fictive example) 20
5 Heterogeneity classification 29
6 Candidates for empirical prior distributions in the historical data (fictive example) . 32
7 Models for the prediction of the new control mean 36
8 Posterior Markov Chain Monte Carlo draws in the historical data (fictive example) 37
9 Bayes factor classification 50
11 External meta data (bending example) 65
12 Variables in the external data (bending example) 65
13 Prior distributions for the historical data (bending example) 69
14 Prior distributions for the historical data (micro-CT example) 70
15 Prior distributions compared in sensitivity analyses 73

16 Internal historical data (bending example) 79
17 Posterior summary of the historical data model (bending example) 80
18 Parametric approximation by maximum likelihood methods (bending example) . . 80
19 Sample size result (bending example) 83
20 Sensitivity analyses (bending example) 83
21 Examination of additional design quality criteria (bending example) 84
22 Design under heteroscedasticity (bending example) 85
23 Internal historical data (micro-CT example) 86
24 Posterior summary of the historical data model (micro-CT example) 87
25 Parametric approximation by maximum likelihood methods (micro-CT example) . 88
26 Sample size result (micro-CT example) 89
27 Sensitivity analyses (micro-CT example) 90
28 Examination of additional design quality criteria (micro-CT example) 90
29 Design under heteroscedasticity (micro-CT example) 91

30 Summary of the historical data (fictive example) 93
31 Posterior summary of the historical data model (fictive example) 94
32 Parametric approximation by maximum likelihood methods (fictive example) . . . 95
33 Bayesian sample size result (fictive example) 96
34 Prior-data conflicts for the means and the standard deviation (fictive example) . . 98

List of Examples

1 Classical sample size determination 12
2 Historical control data 19
3 Conjugate prior distributions in a partly Bayesian analysis 24
4 Bayesian analysis using Markov Chain Monte Carlo computation 27
5 Prior distributions for a Bayesian model of the historical data 32
6 Prior predictive checks for the Bayesian model of the historical data 33
7 Predictive distribution for the new control animals 37
8 Parametric approximation of the predictions 40
9 Bayesian credible interval 45
10 Bayes factor 51
11 Design simulation 54

List of Summaries

1 Predictive prior distributions 41
2 Frequentist vs. Bayesian inference 52
3 Sample size determination by simulation based design analysis 62

List of Main Results

1 Sample sizes for the bending stiffness example 82
2 Sample sizes for the micro-CT example 89
3 Sample sizes for the fictive meta-analysis example 95

Publications as part of this thesis

This thesis is formulated as a monograph that takes up and extends the ideas of designing animal experiments by Bayesian meta-analytic predictive approaches, that I published as preprint on the arXiv server as

> Unseld, T. Designing translational animal experiments by Bayesian MAP approaches. 2023. URL: https://arxiv.org/abs/2301.05572.

on January 13, 2023 under a CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/) [1]. In that contribution, I proposed a concept for using the predictions from a Bayesian meta-analysis model of historical individual-level animal data for the design of a new animal experiment. In the present thesis I will illustrate alternative prediction models, reflecting possible relations between the historical animal data and the new animal data, and develop a systematic workflow for sample size planning in translational animal experiments. The exemplary historical animal data and the application example, that were used in my previous contribution, will be reanalyzed in this thesis and will be extended by a second experimental outcome and an additional application example. This thesis recapitulates concepts of my previous contribution to an extent which is necessary to read this thesis as a coherent, stand-alone work.

1 Introduction

Translational research constitutes a key element in the development of new methods and therapies for pathological phenotypes in human medicine. Situated in the late stage of preclinical research, it aims to translate findings from basic laboratory preclinical research into clinical applications by investigating potential treatments in animals. These investigations are, however, discussed controversially because they often imply injections, surgical interventions, the induction of diseases, or other harm to the animals. Consequently, attempts to find alternatives to animal experiments and to set standards have been developed, like the principles of Replacement, Reduction and Refinement (3R) from Russel and Burch [2]. The 3R principles are ethical guidelines that intend to lower the suffering in animal experiments to a minimum and to raise their standards to a level that is similar to human studies. The 3R principles serve as a basis for the strict legal guidelines concerning the planning and conduct of animal experiments in the European Union, as outlined in the directive 2010/63/EU [3], and in Germany in the Animal Welfare Act ("Tierschutzgesetz") [4] as federal law. Furthermore, the German administrative regulation of animal experiments includes a duty of disclosure and an authorization requirement by an independent Animal Welfare Act Commission ("Tierschutzbeauftragte") with expertise in biomedical and laboratory animal science [5]. Researchers can apply for this authorization by submitting an animal experiment proposal. The general application forms vary, depending on the federal state, but generally, they encompass a description of the analysis methods, including the choice of any statistical hypothesis tests of interest, and a justification of the proposed sample size, which may be based on the choice of the statistical test and on assumptions on the future experiment [6].

Statistical hypothesis tests are constructed to test null hypotheses for parameters of interest in the statistical models of the animal data. For example, it may be of interest to make a decision on whether the knockout of a gene, that seemed relevant for a pathological phenotype in a preceding genome study, ameliorates the pathological condition in vivo. In this case, a new animal experiment could be set up, in which control animals with the pathological phenotype of interest are compared to experimental animals, in which the gene is knocked out, and the null hypothesis of equal group means is tested. The choice of the statistical test is then made in light of assumptions on the future phenotypic measurements and sample sizes can be set up with the attempt to reach a sufficient statistical power of the hypothesis test (as probability of the test correctly

rejecting the null hypothesis in case of truly present effects), while limiting the type I error rate (as probability of the test falsely rejecting the null hypothesis) by a significance level.

What limits the value of formal sample size calculation in practice is that the group sizes in translational animal experiments are typically kept small for animal welfare purposes [7, 8]. This raises challenges in the experimental design. Firstly, smaller sample sizes generally require the presence of larger effect sizes (with otherwise fixed design parameters) to make claims of significant results [9–11]. As a consequence, experimental treatments that come with smaller effects are likely to be missed and if a truly small effect turns out to be significant in a single experiment, it will likely be overestimated [12]. Such claims of significant findings, that turn out to have happened only by chance and cannot be confirmed in subsequent analysis with new data, have contributed to a reproducibility crisis in animal research [13–16]. Secondly, the distributional assumptions for the sample size calculations of future experiments are often based on the observed measurements from previous historical experiments [17]. There, small sample sizes of the historical experiments are problematic, because the estimates are characterized by high uncertainties. These uncertainties are, in general, not adequately reflected with classical frequentist methods and may lead to unreliable sample size estimates for the new experiment. Planning animal experiments with more animals than necessary is both uneconomic and unethical for animal welfare purposes while planning animal experiments with fewer animals than necessary leads to underpowered experiments [8], meaning that the hypothesis test does not actually have the attempted power in the new experiment. Underpowered experiments are problematic because they not only decrease the probability of deciding against the null hypothesis when the hypothesis is, in fact, false (power) but also increase the probability of deciding against the null hypothesis when the hypothesis is, in fact, true (type I error) [18].

Several suggestions have been made to improve poor experimental designs in animal research [6, 15, 16, 19–25]. From a statistical perspective, firstly, systematic reviews and meta-analysis have been proposed as tools to gather and jointly analyze all relevant information for a specific research question [26–30]. A benefit of these approaches over the isolated analysis of single experiments is that the precision of the estimates can be increased, which in turn increases statistical power. In addition, meta-analyses enable the investigation of heterogeneity in different groups and experiments and the generation of new hypotheses [31]. Secondly, Bayesian approaches have been suggested as alternatives to frequentist analysis methods [8, 25, 32–36]. In a Bayesian analysis, a priori knowledge is formally incorporated as a probability distribution into the analysis of new data. The analysis results in an updated posterior distribution, which is then used to answer the medical question of interest. By the incorporation of a priori knowledge in the design and analysis of the new experiment, the application of Bayesian methods has the

potential to decrease the necessary sample size. In addition, in the challenging situation a of few, small historical animal experiments, Bayesian meta-analyses have shown to be superior to frequentist meta-analyses [37].

Despite their appealing properties, there are few real-world examples where Bayesian methods are applied for planning and analyzing preclinical animal experiments [8, 33, 34]. A challenge in the application of Bayesian methods is the formal specification of a prior distribution. The choice of a bad prior distribution can lead to wrong posterior conclusions concerning the medical question. Especially in the situation of animal experiments with small sample sizes, the choice of the prior distribution can have a major impact on the conclusions [38]. To translate the information from historical data into prior distributions for a new experiment, assumptions on the relatedness of the historical and the new experiment have to be made [32, Section 3.16]. The strongest assumption is that the distribution parameters in the new and historical experiments are all identical. In this case, common parameters can be estimated for all available historical experiments and their posterior distributions can be directly used as prior distributions for the new experiment. This strong assumption can only be justified under strictly controlled experimental conditions [39]. A less strict assumption is modeled in an equal-but-discounted approach [32, Section 3.16], where the historical information is discounted as compared to the information from the new experiment. Such an approach has been suggested for the control groups of animal experiments by Bonapersona et al. [8]. The historical data may be characterized by a common pathological phenotype, which is measured by the same or similar outcomes, and investigated under several experimental conditions. The similarity of the animals is then expected to be greater for the control animals than for the experimental animals, since the experimental condition may vary between the experiments, whereas the pathological phenotype defines the control animals in all the experiments. To regulate the impact of individual experiments on the posterior distributions, Bonapersona et al. [8] put weights on the historical experiments. A Limitation of this approach is, however, that the choice of these weights may be challenging for the experiment's researcher and may be criticized because of its subjective nature [40].

An alternative assumption on the relatedness of the historical and the new data is that the data are similar or "exchangeable" [41]. This assumption is represented in a random-effects meta-analysis of the historical data [39, 42–45]. In the context of clinical trials, a meta-analytic-predictive (MAP) approach has been suggested for the mean parameter in normally distributed data by Neuenschwander et al. [39] and for the variance parameter by Schmidli et al. [46]. A limitation, however, of the approach by Neuenschwander et al. [39] is that they model the residual standard deviation (SD) (as the square root of the variance) by a fixed value, instead of a random variable with a prior distribution. Using a fixed parameter value is undesirable from a

Bayesian perspective, since it ignores the uncertainty about the true parameter value and may lead to sample sizes that are too small or to scientific claims that are overconfident [47]. Schmidli et al. [46] modeled the new experiment's residual SD as a random variable and used the historical data to assign it an informative prior distribution. However, for the mean parameter of the new experiment, Schmidli et al. [46] used an uninformative uniform prior distribution, without consideration of historical evidence. Moreover, both the applications of the equal-but-discounted approach by Bonapersona et al. [8] and the exchangeability approach by Neuenschwander et al. and Schmidli et al. [39, 46] used aggregated historical data (as estimates of the mean and residual SD) instead of individual-level data. This can be disadvantageous in the situation of small sample sizes in the historical data, where the estimates may not be very good or when the models are not flexible enough to model experiment-specific characteristics. Instead, if the data are available on an individual-animal level, an individual animal data meta-analysis (IADMA) (also called individual patient data meta-analysis (IPDMA) in the context of human studies) is considered as "gold standard of review" [48], since it is more informative and flexible and better suited to model subgroup effects.

Furthermore, Bayesian methods, that are developed for the application in human clinical trials, typically need to be adapted to the characteristics of animal experiments. Walley et al. [33] have discussed the approach of Neuenschwander et al. [39] in the context of animal experiments. Walley et al. [33] suggested using historical control animals in relation to their role in the future experiment. Furthermore, they have illustrated their thoughts in two real case examples. However, they have not addressed the issue of sample size calculations for animal experiments. Bayesian frameworks for sample size calculation in clinical trials have been illustrated by Schmidli et al. [46] and Wandel et al. [49]. Furthermore, Bonapersona et al. [8] have applied power simulations featuring prior distribution for the mean and residual SD in the control group as power prior distributions [50] that are based on aggregated historical animal data. Moreover, Novick and Zhang [25] have illustrated power simulations in Bayesian models that incorporate random treatment interaction effects of the animal experiment. However, Novick and Zhang [25] did not outline systematic concepts for setting prior distributions for these calculations.

All the approaches of Neuenschwander et al. [39], Schmidli et al. [46], Wandel et al. [49], Novick and Zhang [25] and Bonapersona et al. [8] have used statistical hypothesis that are based on whether the treatment effect, that is assumed under the null hypothesis (often zero), lies inside the posterior credible interval of the effect. Credible intervals are Bayesian versions of frequentist confidence intervals with slightly different interpretations and can be derived from the treatment effect's posterior distribution to make statements on the probability of statistical hypotheses. Together with other Bayesian indices for hypothesis testing, tests based on the

posterior distribution of an effect have shown preferable properties over frequentist p-values in both theoretical considerations [51, 52] and in simulations [53]. From an interpretational point of view, however, hypothesis tests based on posterior credible intervals have been criticized, because they violate the principle of predictive irrelevance [54]. As an alternative to hypothesis testing based on credible intervals, hypothesis tests based on Bayes factors have been suggested [55]. A Bayes factor is generally defined as the ratio of the probability of the observed data (likelihood) under two competing models. In the context of hypothesis testing, these models may represent the null and alternative hypothesis. The Bayes factor then quantifies the evidence for the null hypothesis relative to the alternative hypothesis and concepts have been suggested for using Bayes factors for design analysis and sample size determination [52, 56, 57].

In this thesis, as an extension of the approaches by of Bonapersona et al. [8], Neuenschwander et al. [39] and Schmidli et al. [46], a systematic approach for sample size determination in translational animal experiments using Bayesian analysis and predictions from historical individual-level animal data will be introduced. The proposed approach builds on my previously proposed contribution [1] on designing animal experiments by Bayesian meta-analytic approaches to set up a systematic workflow for sample size determination of a new animal experiment. In my previous contribution, I fitted a Bayesian meta-analysis (hierarchical) model to a combination of individual-level historical animal data consisting of both internal data from the researcher who plans a new experiment, and external data from a phenotypic animal database [58], and used the model to make predictions for the new experiment. In this thesis, I will extend the prediction framework to reflect different assumptions on the relatedness of the historical and the new data. When the historical data stem from the new experiment's researcher, the new animals are considered as additional animals to the historical experiment and predictions will be made under the assumption of full equality of the relevant parameters in the historical and the new data. Instead, when the experimental conditions in the historical and the new data differ, the assumption of full equality will be relaxed to an assumption of similarity (exchangeability) of the experiments [59, Section 5.2]. Furthermore, predictions reflecting covariate effects, like the animal's age, will be considered. Based on the predictions, prior distributions will be derived for both the mean and the residual SD of the new data, thus combining the approaches of Neuenschwander et al. [39] and Schmidli et al. [46] and extending them to a meta-analysis framework with individual-level animal data.

In sensitivity analyses, the sample size obtained from a completely Bayesian model of the new data, with predicted prior distributions for both the mean and SD parameter, will be compared to results from a partly Bayesian analysis, where the residual SD is modeled as a fixed value. In addition, the results will be compared to the results from a Bayesian analysis where the information content of the mean parameter is chosen to match the information content of a single

animal in the new experiment ("unit information prior" [60]). Moreover, the sample size results from hypothesis tests based on the Bayes factor will be compared to the sample sizes with hypothesis tests based on posterior credible intervals. To investigate if the calculated sample sizes are large enough also in the situation where the assumptions and prior distributions are based on the model of the historical data, but the model is biased, various forms of prior-data conflict [8, 38, 61–64] will be investigated. The Bayesian models will be estimated using the software package `brms` [65] in the statistical software language `R` [66]. The `brms` package offers a flexible Bayesian regression modeling framework using modern Markov Chain Monte Carlo (MCMC) computation. With this flexibility in the regression models, the methodology, that is here outlined for the special case of a new experiment comparing two independent and normally distributed treatment groups, can easily be extended to more complex designs or different distributions. The theoretical considerations will be accompanied by a real-world animal experiment with two experimental outcomes (where the historical data for one of the outcomes was used in the analysis of my previous contribution [1]) and by a simulation example, where the sample size results from Bayesian analysis will be compared to the sample sizes results from frequentist tests.

In summary, the main research question in this thesis is:

- What is a systematic approach to the sample size determination for a new animal experiment using predicted prior distributions from historical individual-level animal data in a Bayesian analysis of the new experiment, and can such an approach improve the planning and analysis of new translational animal experiments compared to classical frequentist approaches?

In addition, the secondary research questions in this thesis are:

- What is the effect of the choice of the Bayesian hypothesis test (Bayes factor vs. credible interval based) on the power and sample size results?
- What effect does the model of the standard deviation (SD) have on the sample size results (random variable vs. fixed value)?
- What are the effects of deviations of the new experiment's true parameters from the central tendencies of the prior distributions (prior-data conflicts)?
- Does a Bayesian predictive approach to the planning and analysis of translational animal experiments outperform classical frequentist approaches with respect to additional design criteria to statistical power?

After the introduction in this chapter, Chapter 2 discusses the problems that are associated with sample size calculations in a classical frequentist framework, and introduces Bayesian ap-

proaches for planning and analyzing new experiments based on historical data as its alternative. Meta-analyses are introduced as a method to summarize the historical data from multiple historical experiments and to get a better planning foundation for the new experiment. For the investigation of experimental designs with unequal treatment group variances, a distributional Bayesian model is suggested. The considerations are applied to a real-world application example and to a fictive multi-experiment example in Chapter 3. In Chapter 4, the methods and results from Chapter 3 are discussed, possible extensions are addressed and conclusions on the research questions of this thesis are formulated. The thesis finishes with a summary of the research questions, the methods, the results, their discussion, and the conclusions in Chapter 5.

2 Material and Methods

The following chapter starts by briefly reviewing the components of classical frequentist sample size estimation and by illustrating the problems that are associated with a sample size estimation based on a single small historical animal experiment. In the second part of the chapter, meta-analytic and Bayesian methods are introduced as ways to summarize historical data from multiple experiments and to derive predictions for the new experiment with the aim of increasing the information content of the data for both the planning and analysis of the new experiment. Bayesian inference works by combining a priori knowledge about the parameters in a data (likelihood) model (as the mean and residual standard deviation (SD) in normally distributed data) with the observed data to derive updated posterior distributions. Bayesian models can flexibly be estimated by Markov Chain Monte Carlo (MCMC) methods, which provide non-parametric approximations of the posterior distributions and are derived as series of correlated parameter values. Different methods for making predictions based on the posterior MCMC draws from a Bayesian model of individual-level historical animal data are introduced, representing different assumptions on the relatedness of the historical and new animal data. The predictions are transformed into parametric distributions and included as prior distributions in the planning and analysis of the new experiment. In the third part of the chapter, a concept for performing sample size calculation with Bayesian methods, using the predictive prior distributions from the historical data, is suggested. To this end, Bayesian hypothesis tests are defined in terms of the Bayes factor indicating at least moderate evidence for the alternative hypothesis or as Bayesian posterior credible interval excluding the value under the null hypothesis. In the fourth part, the planning of a real-world animal experiment in translational research is introduced, that addresses bone quality in mice under pathological conditions. Additionally, data from multiple, similar, small experiments are simulated, as a fictive application example of a Bayesian predictive approach in a meta-analysis model of individual-level animal data. At the end of this chapter, software and tools for fitting the Bayesian models in practice using the statistical software `R` are described and details on the fitted models are provided.

2.1 Classical sample size determination in animal experiments

According to the ARRIVE guidelines [67, 68] (Animal Research: Reporting of In Vivo Experiments) (Essential 10 item 2a), the total number of animals (or experimental units) that are used in an animal experiment should be specified, in addition to the allocation ratio to each of the experiment's group. Furthermore, the decision on the sample size should be explained and, if applicable, details on the sample size calculation should be given (Essential 10 item 2b). Ultimately, the sample size should be large enough to detect truly present treatment effects with a high probability (statistical "power", often 80%), while limiting the probability of false claims ("type I error rate" or false positive rate (FPR)) by a predefined significance level α (often $\alpha = 5\%$). The concepts of "power" and "type I error rate" depend on the decision rule used to answer the researcher's biological question. What is the best decision rule depends on the distribution of the data in the new experiment. Here, it is assumed that the interest of the experimental researcher lies in the comparison of an experimental treatment with a control group. This situation often arises in translational research when both the experimental and the control group are set up with an animal model of a pathological phenotype of interest and during the conduct of the animal experiment, a treatment is applied to the experimental group. If the treatment in the experimental group ameliorates the pathological conditions, then it may serve as a candidate for the development of therapies in clinical trials in humans.

For example, the pathological phenotype may be amyotrophic lateral sclerosis (ALS) which is a motor neuron disease that is characterized by selective degeneration of the lower motor neurons and can be investigated in a superoxide dismutase 1 (SOD1) animal model, where the animals have a mutation in the SOD1 enzyme. Another pathological phenotype of interest may be postmenopausal osteoporosis, which can be investigated in ovariectomized mice. The phenotype of postmenopausal osteoporosis will be the pathological phenotype in the later application example in Chapter 3. The impact of the disease is measured, for example, by the mean protein concentration of the cytokine-like interleukin 4 and the grip strength as experimental outcomes. An experimental treatment may be the knockout of a specific gene that was identified in genome-wide association studies. The knockout in the pathological mouse model shall provide information about the biological pathways that are associated with the gene function and if there is potential to ameliorate or cure the pathological condition in vivo. For example, the influence of hormones like the neuropeptide FF (where F stands for the amino acid phenylalanine), which regulates neural activity in the hypothalamus, could be investigated as an experimental condition with the attempt to better understand the ALS disease [69] or the supplementation with melanin-concentrating hormone neurons could be investigated with the aim to reduce the weight loss that is associated with ALS [70].

The sample size planning procedure in this thesis will be illustrated in the commonly used model of two independent treatment groups with normally distributed measurements as

$$y_{ik} = \theta_i + \epsilon_{ik}, \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma_i^2), \quad k = 1, \ldots, n_i, \; i = C, E, \tag{1}$$

for animal $k = 1, \ldots, n_i$ in treatment group $i = C, E$ (control (C), experimental (E)). The terms ϵ_{ik}, which are independent and identically distributed within each group, are the random errors, and $\theta_i = \mathrm{E}(y_i)$ are the group-specific means. More complex designs and alternative distributions are discussed in Section 4.1.2.

2.1.1 Concepts of frequentist hypothesis testing

The research question, that shall be answered in the new experiment, is if the (absolute) difference $\delta = \theta_E - \theta_C$ in the treatment group means is different from the null value $\delta_0 = 0$. This research question is formulated as a hypothesis pair of a null and an alternative hypothesis:

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta \neq 0 \tag{2}$$

Using a statistical test, the researcher wants to find out if H_0 can be rejected based on the observed data. In the case of two independent, normally distributed groups from equation (1), the two-sample test is used for the situation of equal variances (homoscedasticity) with test statistic

$$T = \frac{\overline{y}_E - \overline{y}_C}{S_N \sqrt{\frac{1}{n_E} + \frac{1}{n_C}}}. \tag{3}$$

The test statistic is calculated from the arithmetic group means and the empirical variances:

$$\overline{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik} \tag{4}$$

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_i)^2, \tag{5}$$

for group $i = C, E$. Then, $S_N^2 = \frac{1}{N-2}((n_C - 1)\hat{\sigma}_C^2 + (n_E - 1)\hat{\sigma}_E^2)$, with $N = n_C + n_E$, is the pooled variance estimator and the absolute value of the test statistic is compared to the critical value $t_{N-2,1-\frac{\alpha}{2}}$ as the $1 - \frac{\alpha}{2}$ quantile of the student t-distribution with $N - 2$ degrees of freedom [71, Chapter 3]. The corresponding two-sample t-test rejects H_0 at significance level α if

$|T| > t_{N-2,1-\frac{\alpha}{2}}$. If there are signs of unequal variances (heteroscedasticity) in the data, Welch's t-test is used instead with the test statistic

$$T_{\text{Welch}} = \frac{\overline{y}_E - \overline{y}_C}{\sqrt{\frac{\hat{\sigma}_C^2}{n_C} + \frac{\hat{\sigma}_E^2}{n_E}}}. \tag{6}$$

In the Welch test, the absolute value of T_{Welch} in the data is compared to the critical value $t_{\tilde{\nu};1-\frac{\alpha}{2}}$ as the quantile of the t-distribution with modified degrees of freedom $\tilde{\nu}$ (corresponding to the approximate distribution of the test statistic under H_0 [72, 73]). H_0 is then rejected with the Welch test at significance level α if $|T_{\text{Welch}}| > t_{\tilde{\nu};1-\frac{\alpha}{2}}$.

2.1.2 Problems in the design and analysis of translational animal experiments

Given a statistical test for a predefined significance level α, the sample size for a new experiment is typically chosen with the goal to reach at least a certain statistical power. The *power of the test* is the probability that the test rejects H_0 if the data stem from the model under H_1 ("H_1 is true"). If the test does not reject H_0, although H_1 is true, a type II error is made. If the test rejects H_0, although the data stem from the model under H_0 ("H_0 is true"), a type I error is made. Given conventional values for α, which limits the type I error rate, and a target value for the power as $1-\beta$ (where β denotes the probability of type II errors), the sample size $N^* = n_C^* + n_E^*$ in the new experiment is calculated by optimization of power equations as the minimal total number $N = n_C + n_E$ that gives the desired power. For the t-test and the Welch test, the critical value themselves ($t_{N-2,1-\frac{\alpha}{2}}$ and $t_{\tilde{\nu};1-\frac{\alpha}{2}}$) depend on the sample size through their degrees of freedom. As a consequence, no analytical solution is available for the power equation. In practice, the sample sizes are thus found by iterative search procedures or by approximations [74, Section 3.3]. For example, for a large sample size, the quantiles of the t-distribution can be approximated by the quantiles of a normal distribution, in which case the total sample size can (under the assumption of equal group SDs $\sigma_C = \sigma_E = \sigma$) be calculated as

$$N = \frac{(1+a)^2}{a}\left(\frac{u_{1-\alpha/2} + u_{1-\beta}}{\delta^*/\sigma}\right)^2,$$

where $u_{1-\alpha/2}$ and $u_{1-\beta}$ are the $1-\alpha/2$ and $1-\beta$ quantiles of the standard normal distribution, δ^* is the "minimal (biologically or clinically) relevant mean difference", i.e., the minimal (absolute) difference in the group means that the researcher wants to be able to detect with the resulting sample size, and $a = \frac{n_E}{n_C}$ is the allocation ratio of the total sample size to the two groups [74, Section 3.2].

The sample sizes in the control and experimental group are then:

$$n_C = \frac{1}{1+a} \cdot N$$
$$n_E = \frac{a}{1+a} \cdot N$$

In most situations (in particular for the two-sample t-test with equal group variances) the total sample size is minimized by allocating half of the total sample size to each group, i.e. by setting $a = 1$ [74, Section 3.3]. In the Welch test situation of unequal variances, the choice of an optimal allocation ratio is more complicated and depends on the ratio of the two variances [75]. Conventional values for α and β are $\alpha = 5\%$ and $\beta = 20\%$ [7, 17]. In practice, δ^* and σ are often estimated from historical data, as illustrated in the following example.

Example 1: Classical sample size determination

For planning a new experiment, the researcher of an animal experiment provides data from a historical experiment where measurements in control animals and experimental animals were made. Under the assumption that the parameters in the new experiment are the same as in the historical experiment, the empirical means and standard deviations (SDs) in the historical data are used as for calculating the sample size calculation in the new experiment.

To illustrate such a scenario, hypothetical historical data are simulated as $y_{Ck} \sim \mathcal{N}(0, 1)$ and $y_{Ek} \sim \mathcal{N}(1.5, 1)$, $k = 1, \ldots, 5$ with true mean difference of $\delta = 1.5$ and true SDs of $\sigma_C = \sigma_E = 1$. A mean of zero and a SD of one in metric data can always be achieved by standardization of the data (subtracting the mean and dividing by the sample SD). In practice, the groups' true means and SDs are unknown and researchers can only draw conclusions about the parameters by calculating estimates from the observed historical data. In a sample of five hypothetical animals per group, estimates for the mean parameters and standard deviations (as the square root of the sample variances) are calculated according to equations (4) and (5). The results as given in Table 1. The observed data only represent a sample of the population for which a researcher wants to make medical claims. This involves random noise in the simulated data and leads to empirical estimates that deviate from the true values. In particular, the following bias occurs in this sample of historical data:

- Bias in the mean difference δ: in the control group, the true mean $\theta_C = 0$ is overestimated by the arithmetic mean $\overline{y}_C = 0.2$. In the experimental group, the true mean

$\theta_E = 1.5$ is underestimated by the arithmetic mean $\overline{y}_E = 0.7$. This leads to the true mean difference $\delta = 1.5$ being underestimated by the difference in the arithmetic means $0.7 - 0.2 = 0.5$.

- Bias in the standard deviations σ_C, σ_E: The true standard deviations $\sigma_C = \sigma_E = 1$ are underestimated by the empirical standard deviation in the historical control group $\hat{\sigma}_C = 0.9$ and overestimated by the empirical standard deviation in the historical experimental group $\hat{\sigma}_E = 1.33$).

Table 1: Summary statistics in a data set of hypothetical historical data with two groups (control (C) and experimental (E)). Mean: arithmetic mean. SD: empirical standard deviation. n: group size.

Group	Mean	SD	n
C	0.20	0.90	5
E	0.70	1.33	5

For planning the new experiment, the researcher uses the difference in means $\hat{\delta} = 0.5$ as an estimate for the biologically or clinically relevant effect δ^* in the new experiment and the empirical SDs as estimates for the residual SDs of the new groups. Since the empirical SDs in the two groups are different, the researcher uses the Welch test (6) instead of the t-test (3). With a significance level of $\alpha = 5\%$, a target prospective power of $1 - \beta = 80\%$ and an equal allocation ratio of the total sample size N to the control group of $a = 1$ (leading to equal group sizes $n = n_C = n_E = \frac{N}{2}$), the researcher gets a sample size per group (calculated in the statistical software `R` [66] with the `power_t_test` function of the `MESS` package [76] and Welch test degrees of freedom) of

$$n = 83.$$

In contrast, with the true parameters $\delta = 1.5$ and $\sigma_C = \sigma_E = 1$ from above, the sample size with the t-test (calculated in the statistical software `R` [66] with the `power_t_test` function of the `MESS` package [76] and classical t-test degrees of freedom) would be $n_{\text{true}} = 9$. From a practical perspective, a calculated sample size per group of $n = 83$ is typically too large to be accepted by an animal welfare committee reviewing the animal proposal, especially when multiple groups and comparisons are involved increasing the sample size multiplicatively [7]. To increase the chance of the animal experiment proposal being accepted, the researcher may look for different historical data, where a larger difference in

the group means was observed. In Table 2 the estimates of a second historical experiment (sampled from the same distributions as the first experiment) the estimates are given.

Table 2: Summary statistics in a second data set of hypothetical historical data with two groups (control (C) and experimental (E)). Mean: arithmetic mean. SD: empirical standard deviation. n: group size.

Group	Mean	SD	n
C	0.20	0.73	5
E	2.20	0.81	5

In this experiment the true mean difference δ is overestimated and both residual SDs are underestimated, leading (with the Welch test (6) in the `power_t_test` function in R [66] and the same values for the prospective power and the type I error rate as before) to a small number of required animals as

$$n = 4.$$

Since this smaller number is more likely to be accepted by animal welfare committees, the researcher may decide to submit their proposal with this sample size estimate.

Example 1 illustrates common problems of sample size calculation based on historical data from a single small historical experiment [7]. In particular, problems are connected to the following facts:

1. The group sizes of the animal experiments are typically kept very small (ten or less [8]), leading to
   (a) imprecise estimates of the model parameters (mean and residual SD) and in turn to bias in the sample size estimate
   (b) attempts to justify a target sample size by selection of historical experiments that provide convenient estimates
2. It is assumed that the true means in the historical control and experimental groups are the same as in the new experiment. However, in reality, the situation may be more complicated and this assumption may be unrealistic due to heterogeneity in animal batches, laboratories, research personnel, or other factors that vary in the new experiment over the historical experiment. This heterogeneity can lead to yet greater deviations of the historical mean estimates from the true means in the new experiment and must be reflected in the esti-

mation, as will be shown in the next section. Moreover, for the experimental group, it may be inadequate to assume that the treatment effect of the historical experiment coincides with the new treatment. This is because, typically, in a new experiment the investigated experimental treatment changes, and at the early stage of translational research, little is known about its actual effects.

In the following sections, it is explained how these problems can be tackled by using:

(a) Meta-analytic methods to combine historical control data and make predictions for a new animal experiment, with the aim of obtaining more reliable estimates and a better planning basis while reflecting the relationship between the historical and the new experiments

(b) Bayesian methods to incorporate the predictions from the model of the historical data as prior distributions into the analysis of the new experiment, with the aim of reducing the sample size in the new experiment while reflecting the uncertainty in the predictions

2.2 Bayesian predictive approaches

The aim in this part of experimental planning is to fit a (meta-analysis) model to available control data, that are assumed to be related to the control data in the new experiment, and to use the model for making predictions for the mean and residual standard deviation in the control group of the new experiment. These predictions will then be used as prior distributions and parameter estimates in the second part of the experimental planning, as introduced in Section 2.3. The procedure of fitting Bayesian models to available and relevant control data and using their posterior distributions to make predictions for planning and analyzing the new experiment will be explained in more detail in the following sections.

2.2.1 Analysis of historical control data

Due to the complexity of pathological phenotypes, relevant candidates for treatments in humans can hardly be discovered in one single translational animal experiment. Instead, pathological phenotypes are usually examined in a series of experiments run over multiple years, where different treatment options are compared to similar control animals [34]. Consequently, when a new experiment is planned in such a series of experiments, the control group in the new experiment can be expected to be similar to the historical control animals and it is reasonable to include historical information in the planning and analysis of the new experiment to get better estimates [8].

In the following section, the first part of the Bayesian predictive sample size planning procedure of this thesis is explained, as schematically displayed in Figure 1.

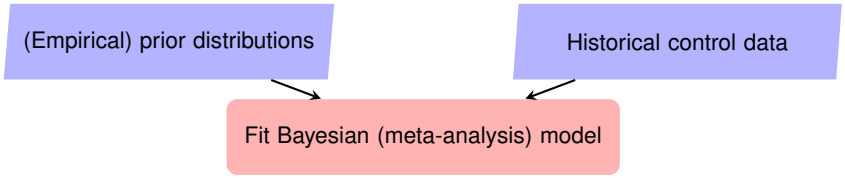


Figure 1: Historical model part of the sample size planning procedure with Bayesian predictive approaches.

Depending on the assumptions on the relatedness of the historical data and the new data, different models for the prediction of the new data are considered. Possible models are presented in Table 3.

Table 3: Assumptions for the relatedness of the new and the historical control parameters (based on [32, Section 3.16, p.90f] and [39, 77]).

Model	Assumption on a parameter in the new control group w.r.t. the corresponding historical control parameters
(a)	Identical as in the historical data
(b)	Differs from the historical parameters by the (fixed) population effect of some covariate
(c)	Identical as the parameter in a specific historical subgroup, where the subgroup parameters vary randomly around a common parameter
(d)	Represents the variation across all the historical subgroup parameters from (c)
(e)	Brand new but related/ similar to the historical parameters by sharing the same model of variation around the common parameter from (c)
(f)	Independent of the historical parameters

In this thesis, the models (a), (b), (e) and (f) are considered for the control group means. Meanwhile, for the residual SD, a common SD is assumed in the historical and new control animals (model (a)). Alternatively, covariate effects or heterogeneity may be as well modeled in the residual SD, e.g. by fitting a distributional model to the historical data, in which predictor terms are specified not only for the mean but also for the residual SD [78]. This topic will be further discussed in Section 2.2.3 and Section 4.1.1. Model (a) is adopted in the common approach of frequentist sample size calculation based on a single historical experiment, like in Example 1. When no other information is available, it may be the best possible assumption. In addition, the model may be reasonable under strictly controlled conditions, where the new experiment and all the other historical experiments are conducted by the same researcher. Then, the new data may be viewed as equivalent data to all the other historical experiments. However, when the parameter in the new control group differs from the parameters in the historical control group, the model

(a) fails to correctly represent the dispersion of the data around the common mean and produces biased predictions.

Model (b) may occur when there are population variables like the animals' age or sex that are known to influence the experimental outcome of interest and whose values are different in the historical and the new experiment. For example, the historical data may consist of measurements in animals of different ages between ten and thirty weeks at the time of the measurement, and the new experiment may consist of animals that are twenty weeks old at the time of the measurement. Then, the effect of age can be estimated from the historical data, and it can be used to make a prediction, given the age of the new animals. If differences are expected to be due to a categorical variable, like the animals' sex, the model of the historical data may either be fitted only to the animals that are of the same sex as in the new experiment, or the variable sex may be included in the prediction for the new experiment. The latter case may be useful to achieve a higher precision in the estimation of the residual variance, in the case where the residual variance is assumed to be the same in both sexes. Otherwise, the exclusion of the other sex may be the better option. The same applies to model (c): since only the information from a specific subgroup g^* is assumed to be relevant for the prediction of the new parameter, the other subgroups may only be useful for the prediction of additional model parameters, but not for the parameter of interest. Model (c) may be used, for example, when the same researcher collects additional data for a specific previous historical experiment [59, Section 5.2].

Like model (a), models (d) and (e) refer to the overall parameter in the historical data for making predictions, but the predicted data will have a wider variation than model (a), since they reflect not only the within-group, but also the between-group variation. Model (d) can be used if the interest lies in a reflection of the variation across all the existing historical subgroup parameters in the prior distribution for the new parameter. Model (e) can be used if the interest lies in building a prior distribution for a new experiment, that has different properties than the historical experiments, but the researcher still expects some similarity or *exchangeability* between the experiments' data [1, 39]. Exchangeability means that there are no systematic differences in the new and the historical data [79, p.10]. This assumption is modeled by a random-effects meta-analysis of the historical data.

Model (f) may be used when the researcher finds, that the available historical data are characterized by properties, that make the historical control data so different from the new control group, that the historical data are considered irrelevant for the prediction of the new data. As stated before, the historical data then may still be useful in adding information to the model of the residual variance. Moreover, while a model of such historical data may not be useful for the definition of a prior distribution for the parameter in the new control group of interest, it may still be relevant

for the prediction of further control groups, that are of secondary interest. For example, the controls of interest may represent untreated animals of a certain pathological phenotype. In addition to those "positive" controls, in which a change as compared to the pre-experimental condition is initiated by the pathological phenotype, further control groups may exist as sham or vehicle controls and wild type animals as "negative" controls, in which no change as compared to the pre-experimental condition is initiated. While positive controls are used for testing the hypothesis of main interest in the new experiment, negative controls are classically included to judge the distribution of the measurement in a healthy population and to assess if there are unmeasured variables that influence the new experiment [80]. Such control groups of secondary interest are further discussed in Section 4.1.1.

Since the sample sizes in translational animal experiments are typically small, a meta-analysis of multiple small experiments may be the best option to increase the overall sample size of the historical data and to get more reliable estimates for the planning and analysis of the new experiment. In the following, the focus thus lies on model (e). In the context of a meta-analysis, it is assumed that there is historical control data from $G > 1$ experiments. The data in each experiment are assumed to be independent and identically distributed as

$$y_{gk} = \theta_g + \epsilon_{gk}, \quad \epsilon_{gk} \sim \mathcal{N}(0, \sigma_g^2) \tag{7}$$

with residuals ϵ_{gk} for animal $k = 1, \ldots, n_g$ and experiment-specific means θ_g in experiment $g = 1, \ldots, G$. Heterogeneity in the means is modeled as normally distributed random deviations from a common mean μ, which may, e.g., occur due to different batches, animal strains, laboratory conditions, or measuring instruments:

$$\theta_g = \mu + \gamma_g, \quad \gamma_g \sim \mathcal{N}(0, \tau^2) \tag{8}$$

Connecting equations (7) and (8), the normal-normal hierarchical model (NNHM)

$$\begin{aligned} y_{gk}|\theta_g, \sigma_g &\sim \mathcal{N}(\theta_g, \sigma_g^2) \\ \theta_g|\mu, \tau &\sim \mathcal{N}(\mu, \tau^2) \end{aligned} \tag{9}$$

defines a meta-analysis of the historical animals $k = 1, \ldots, n_g$ in the experiments $g = 1, \ldots, G$ [59, Section 5.4],[1]. This model is termed *hierarchical* since it connects the individual-level data distribution to the second level distribution of experiment-specific means and it is termed

normal-normal hierarchical model since on both levels a normal assumption is made. If a normal distribution seems inappropriate for modeling the historical data, for example because the data is on the wrong scale or the empirical distribution is skew, a transformation of the data may be considered [81–83]. Alternative models are discussed in Section 4.1.2. In a frequentist framework, this model with normally distributed means $\theta_g \sim \mathcal{N}(\mu, \tau^2)$ is also referred to as *random effects model*.

The meta-analysis or hierarchical model of the historical data provides estimates of the mean, the heterogeneity, and the residual standard deviation parameters in the historical control animals, which can be used to get a better planning basis for the sample size in the new experiment. The meta-analysis model can be fitted both in a frequentist or a Bayesian framework. The advantage of a Bayesian framework is, that the estimates come as probability distributions, which provides a natural way to quantify uncertainties and allows formulating prior distributions for the inclusion of the historical information as prior distribution in a Bayesian analysis of the new experiment. Details of Bayesian analysis will be given in the following section.

Example 2: Historical control data

In this example, there are assumed to be $G = 5$ hypothetical historical experiments addressing the same phenotypic outcome as in a newly planned experiment. The similarity of the historical data is modeled by sampling the means θ_g, $g = 1, \ldots, G$, in the historical data from a common overall mean $\mu_C = 0$. In this example, the heterogeneity is set to $\tau = 0.125 \cdot \sigma_C$ (moderate heterogeneity according to the classification by Spiegelhalter et al. [32, Section 5.7]), where $\sigma_C = 1$ is the within-standard deviation in the historical control animals. The hypothetical data are presented in Table 4. The exemplary data will be used to illustrate the Bayesian predictive approach and sample size planning procedure in Section 2.2.3 and 2.3, under the assumption that the historical and the new data are similar (according to model (e) in Table 3). Furthermore, in Section 3.4, the exemplary data will be complemented with hypothetical experimental data to compare the results from the criticized classical sample size calculation approach (as illustrated in Example 1) to those from the Bayesian predictive approach.

A Bayesian hierarchical model is fit to the historical data. The choice of prior distributions will be discussed in Section 2.2.3 and Example 5. The Bayesian model results in posterior distributions of the mean parameter and the residual standard deviation, representing the combined information from the prior distribution and the data. Details on Bayesian analysis will be provided in the following section. The posterior distributions of the experiment-specific and the overall mean are presented in Figure 2. The meta-analysis "shrinks" the

experiment-specific arithmetic means toward the overall mean. As a consequence, the impact of single extreme observations (outliers) is diminished and the uncertainty in single groups can be reduced by borrowing information from other groups [59, Section 5.6]. The degree of shrinkage for one group depends on its sample size (or the associated standard error) and the between-group variance (heterogeneity) [49, 59, 84–86].

Table 4: Hypothetical historical control data of five experiments. Represented are the experiment-wise summaries of the data as arithmetic means (Mean), empirical standard deviation (SD), and group size (n).

Experiment	Mean	SD	n
1	0.20	0.90	5
2	-0.40	1.04	5
3	-0.08	1.13	7
4	0.51	0.63	10
5	0.08	1.12	10

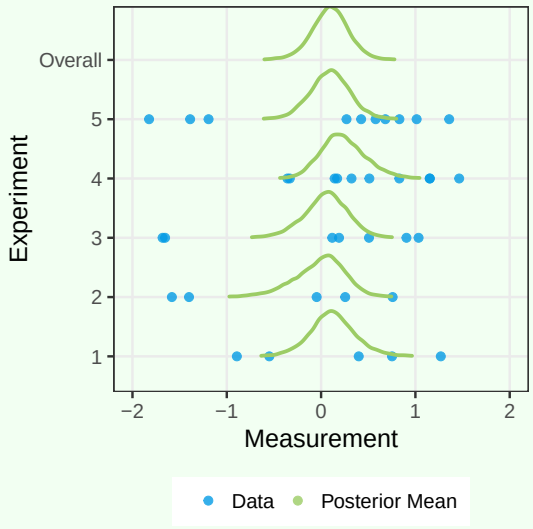


Figure 2: Bayesian estimates (as posterior probability density functions (green lines)) of the experiment-specific means of five fictive historical experiments. The estimates are presented together with the individual-animal measurements (blue dots). In the figure's top row, the estimated overall mean is shown.

2.2.2 Bayesian analysis of animal experiments

The idea of a Bayesian analysis is to formally express a priori information on a model parameter θ (e.g., the expected value in the new experiment's control or experimental group or the difference in the group means δ) in a prior distribution $p(\theta)$ and to update this prior distribution by the data $y_1, \ldots, y_n$ in a new experiment to a posterior distribution using Bayes' rule:

$$p(\theta|y_1, \ldots, y_n) = \frac{p(\theta)p(y_1, \ldots, y_n|\theta)}{p(y_1, \ldots, y_n)}.$$

The distribution $p(y_1, \ldots, y_n)$ of the data is called *likelihood distribution*. It is based on properties of the observed measurements and is here assumed to be from the family of normal distributions as in equation (1). The prior distribution expresses a researcher's assumptions on the data, before considering the new observations. For example, a researcher may expect, that the average of the measurements in normally distributed control data (1) will be with a high probability between minus one and one and most likely be zero. To reflect this assumption, they could assign the mean θ_C from the model (1) a normal distribution that is centered at zero and has most of its probability mass (e.g., 95%) between minus one and one. A normal distribution that approximately fulfills these conditions is given by the prior $\theta_C \sim \mathcal{N}(0, 0.51^2)$.

Problems may arise when the prior distribution is in conflict with information from the data (prior-data conflict, as further discussed in Section 4.1.2). Then, a poor choice of a prior distribution may bias the posterior results and the research conclusions. The impact of the prior on the posterior distribution becomes especially relevant when the sample size in the data is low, as it is typically the case in translational animal experiments [37]. Moreover, the specification of the prior distribution is crucial, when the sample size in the new animal experiments is set up with the goal of reaching a certain prospective power and when this power is defined as a function of the Bayes factor [87] (details on Bayes factors will be given in Section 2.3.1). To derive profound research conclusions, it is therefore important to set prior distributions transparently. A systematic approach for setting prior distributions is to base them on predictions from a model, that is fitted to available historical data [32, Section 3.13, 3.16]. Such an approach will be introduced in the following sections and Bayesian models will be fit in two contexts:

1. Model of the historical data (meta-analysis context with the aim to gather historical information)
2. Model of the new experiment's data (experimental design context with the aim to compute sample sizes for the new experiment)

2.2.3 Choice of prior distributions

In practical applications, a prior distribution should be aligned with the problem at hand and reflect prior beliefs that are specific to the biological application context [87–90]. As a consequence, Bayesian ideals are missed if the same default prior is applied to each animal experiment, independent of the experimental outcome and the available a priori information. The choice of a parameter's prior distribution incorporates assumptions on the distributional family and on the "hyperparameters" as parameters of the distributional family. The choice of a distributional family is made with regard to interpretability, computational efficacy and sensitivity of the posterior distributions to the data. Generally, distributional families with fewer parameters and common distributional families (like the normal distribution) are easier to interpret and discuss with practical researchers of animal experiments than more complex distributional families. When it comes to computational efficacy, some distributional families lead to faster computation in those cases, where they lead to posterior distributions with known probability functions. Regarding the sensitivity to the data, the prior should allow the data to have an impact on the posterior distribution and not be insensitive to the observed data [38, 91]. In the following section, the principles, that were used for the definition of prior distributions for the model of the historical data and the new data, will be introduced. These principles take up the considerations from my previous contribution [1].

Conjugate prior distributions and analytical computation

A popular choice for a parameter's prior distribution is a *conjugate prior*. Given the likelihood distribution of the data, a prior distribution for a model parameter in the likelihood distribution is called conjugate to the likelihood distribution if it results in a posterior parameter distribution, which is from the same distributional family as the prior distribution. This choice has advantages concerning the interpretation and computation of the posterior distributions [91], [59, Section 2.5]. In the case of the normal distribution in the model (1) of the experiment's data, the likelihood distribution depends on multiple model parameters, which are all modeled by their own prior distributions in a "completely" Bayesian model. In contrast, in a partly Bayesian model, only some parameters are modeled as random variables, while others are modeled as fixed values. In this context, *a conditionally conjugate prior* for a parameter is a prior distribution which is conjugate to the likelihood distribution family, conditionally on fixed values of the other model parameters [91]. In the case of normally distributed data as in (1), the family of normal distributions constitutes a conditionally conjugate prior family for the mean parameters, since, for a fixed value of the residual standard deviations, the posterior distribution is again a normal distribution.

More specifically, the normally distributed prior distributions of the treatment group means

$$\theta_i \sim \mathcal{N}(m_i, s_i^2), \quad i = C, E, \tag{10}$$

with mean m_i and standard deviation s_i result in normally distributed posterior distributions [59, Section 2.5]

$$\theta_i | y_{i1}, \ldots, y_{in_i} \sim \mathcal{N}(m_i^*, (s_i^*)^2), \quad i = C, E, \tag{11}$$

where the posterior mean and standard deviation are given as

$$m_i^* = (s_i^*)^2 (\frac{1}{s_i^2} m_i + \frac{n_i}{\sigma_i^2} \overline{y}_i) \tag{12}$$

$$s_i^* = \sqrt{\frac{1}{\frac{1}{s_i^2} + \frac{n_i}{\sigma_i^2}}}, \quad i = C, E. \tag{13}$$

With this choice, the prior distributions can be interpreted as equivalent data. More specifically, the prior means θ_C and θ_E can be thought of as estimates of additional data (to the observed data in the new experiment), that have sample means m_C and m_E and are estimated with total precision (as inverse of the variances) $1/s_C^2$ and $1/s_E^2$. The mean parameters m_i^*, $i = C, E$ from the posterior distributions can then be interpreted as a weighted average of the prior means m_i and the data means $\overline{y}_i$ with weights that depend on the precision of the prior distributions and the sample means [59, Section 2.5]. The choice of the prior distribution's (hyper)parameters characterizes the distribution, like the central tendency and the spread of the distribution. The precision of the prior distribution (as $1/s_i^2$, $i = C, E$) determines its information content and influence on the posterior distribution in relation to the influence of the observed data. A prior distribution can be categorized according to its impact on the posterior distribution to be either non-informative, weakly informative, or informative [59, Chapter 2]. This categorization has to be interpreted as a relative impact in relation to the impact of the data. The higher the precision of the prior is over the precision of data, the more informative the prior is for the posterior distribution. A non-informative prior has a very flat probability density function and is intended to let the information for the posterior distribution come from the data alone. Non-informative prior distributions typically lead to analysis results that are similar to results from frequentist analyses and are a good option when there is enough evidence in the data [59, Section 2.8].

To exploit the potential of Bayesian analysis and especially in the scenario of small sample sizes in animal experiments and when using Bayes factors for hypothesis testing, context-specific weakly informative or informative prior distributions are preferred over non-informative prior dis-

tributions [87–91]. For example, if the prior standard deviation s_C of the control mean θ_C in (10) is chosen to coincide with the standard deviation σ_C of the data and $n_C = 1$ new animal will be observed, then the prior mean m_C and the arithmetic data mean $\overline{y}_C$ (consisting of the single observation) have the same impact or weight in the posterior mean (12). If, with the same prior, $n_C = 10$ new animal will be observed, then the weight of the arithmetic mean $\overline{y}_C$ (consisting now of ten observations) will be ten times the weight of the prior mean m_C. Such a prior distribution, that has the information content (precision) of one single typical observation, is called *unit information prior (UIP)* [60, 92]. a UIP is recommended as a starting point for the situation of no other a priori information [60, 92]. With more precise prior information, the SD of the prior may be decreased to derive a more informative prior, that has a larger impact relative to the data.

In a completely Bayesian analysis, both the mean and the SD of the likelihood distribution are modeled as random variables (instead of a fixed value of the SD in the partly Bayesian analysis). There, the conjugate prior family is the normal-Gamma family [93, p.169]. Distributions in this family are two-dimensional, with the mean in the first dimension and the variance in the second dimension (with details given by DeGroot [93, p.169] and Gelman [59, Section 3.3]). Since this two-dimensional distribution is more difficult to implement as a prior distribution and since it implies more assumptions in terms of the correlation structures of the mean and residual SD, in this thesis, conjugacy of the family is only considered for one-dimensional distributions and the prior distributions for the mean and SD are modeled to be independent, as in most practical applications [46, 94, 95]. Conditionally on the SD being fixed, the normal prior is still conditionally conjugate, although the marginal distribution (averaged over the SD distribution) is not.

Example 3: Conjugate prior distributions in a partly Bayesian analysis

In the bottom line of Figure 3 hypothetical data of ten animals in two groups (experimental (E) and control (C)) of a hypothetical new experiment are shown. The data were simulated as $y_{ik} \sim \mathcal{N}(\theta_i, \sigma_i^2)$ for animal $k = 1, \ldots, 10$ in group $i = C, E$ with true parameters $\sigma_C = \sigma_E = 1$ and $\theta_C = 0$, $\theta_E = 0.8$, leading to a true mean difference of $\delta = 0.8$. The family of normal distributions consists of two parameters, the mean and the variance parameter (or its square root, the SD), which are here denoted as θ_i and σ_i^2, $i = C, E$. In a completely Bayesian analysis, prior distributions are formalized for both the mean and the SD. In a partly Bayesian analysis, prior distributions are formalized for the mean only, as the parameter of primary interest, while the SD, as the parameter of secondary interest, is set to a fixed value. Conditional on fixed values of the residual SDs σ_C and σ_E, the prior distributions $\theta_i \sim \mathcal{N}(m_i, s_i^2)$, $i = C, E$, from the normal distribution family lead to posterior distributions that are again normally distributed and (conjugate prior distributions). Setting

$s_i, i = C, E$, for the mean parameters' prior distributions to the same value as the SD of the data leads to unit information priors (UIPs) that are designed to have the same information content as a single animal in the observed data. With this choice, the effective sample size (ESS) of the prior (indicating how many observations in the new data the prior is worth, as explained in Section 2.2.3) is one. The posterior means for θ_i are then calculated as weighted averages of the prior means m_i, weighted by one, and the empirical means $\overline{y}_i$, weighted by the sample size of $n_C = n_E = 10$, so $m_i^* = (1 \cdot 0 + 10 \cdot \overline{y}_i)/11$. The posterior SDs for θ_i are obtained by diving the common SD of $\sigma_i = s_i = 1$ of the data and the prior through the square root of the sum of the prior ESS plus the sample size of the data ($s_i^* = 1/\sqrt{1 + 10}$). The prior mean $m_i = 0$ and the empirical means $\overline{y}_C \approx -0.159$ and $\overline{y}_E \approx 0.640$ then result in posterior distributions of approximately $\mathcal{N}(-0.145, 0.0.091)$ for the control group and $\mathcal{N}(0.582, 0.091)$ for the experimental group.

The prior distributions and the posterior distributions are shown together with the estimated probability density of the data in Figure 3. The choice of a priori equal means reflects the conservative belief that, before observing the data, the group means are expected to be equal on average and only enough evidence in the data can lead to a posterior mean different from zero. What is "enough" evidence can be judged in terms of a hypothesis test (as further elaborated in Section 2.2.2).

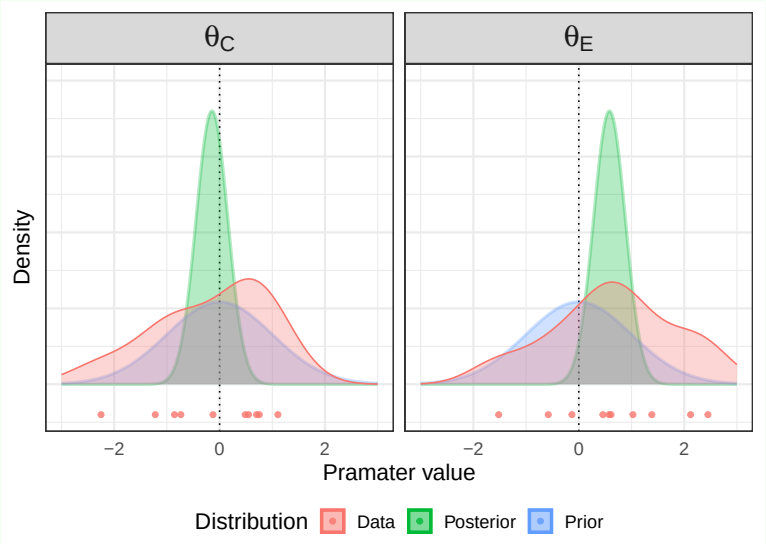


Figure 3: Probability densities of the prior and posterior distributions of the group means θ_C and θ_E and kernel density estimate of simulated data of ten hypothetical animals in two groups (experimental (E) and control (C)).

Markov Chain Monte Carlo computation

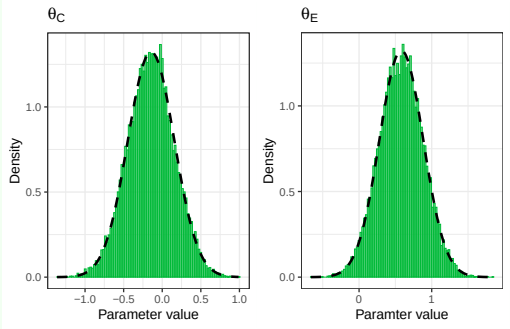
In the partly Bayesian analysis of the previous section, it was assumed that the residual standard deviations σ_C and σ_E are known, and prior distributions were specified only for the mean parameters θ_C and θ_E. In reality, the residual standard deviations are unknown and have to be estimated from available data. In frequentist analysis, the standard deviations are estimated from the data, as the square root of the empirical variances (5). These point-wise estimates could be plugged into equation (12) and (13) to calculate the posterior hyperparameters as before. From a Bayesian perspective, however, fixed estimates are considered overly simplistic since they don't represent the uncertainty in the parameter estimation [59]. Instead, the posterior distributions for the means θ_C and θ_E are better calculated as averages over the full range of the prior distributions for σ_C and σ_E. This is the idea of a completely Bayesian analysis where all model parameters (means, standard deviations) are modeled as random variables. On the one hand, this leads to better representation of uncertainties [47, 59]. On the other hand, it complicates the estimation of the posterior distribution as compared to modeling only one parameter as a random variable, since the higher dimensional joint distributions of all model parameters have to be considered. The calculation becomes even more complicated when the model includes additional parameters, like the animals' ages, and these parameters have prior distributions as well. As a consequence, the posterior density functions in a Bayesian analysis can typically not be expressed in analytic form [59]. Numerical computation using Markov Chain Monte Carlo (MCMC) methods [96, 97] is a common alternative.

A Monte Carlo (MC) algorithm is a technique to calculate functions approximately, by simulation and numerical methods [96]. This involves the iterative generation of simulation draws of the model parameters. The sequence of these auto-correlated draws represents a Markov chain, which is a sequence of random numbers for which the distribution of the next simulated value depends only on the current value, but not on the previous values. Under certain assumptions, the Markov chain converges towards a stationary distribution. The stationary distribution is chosen to correspond to the posterior distribution of interest. According to MC theory, the approximation of the target distribution gets better in each iteration step. To eliminate the influence of the chains' starting values, a burn-in interval is defined, and only the simulation draws after this interval are kept. The remainder of the iterations are discarded as "warm-up" draws. Furthermore, multiple MCMC chains are simulated, to judge their convergence against a common distribution [59]. In this thesis, Markov Chain Monte Carlo (MCMC) methods are used to calculate the posterior distributions in the historical and new experiment and to test hypotheses as described in Section 2.2.2.

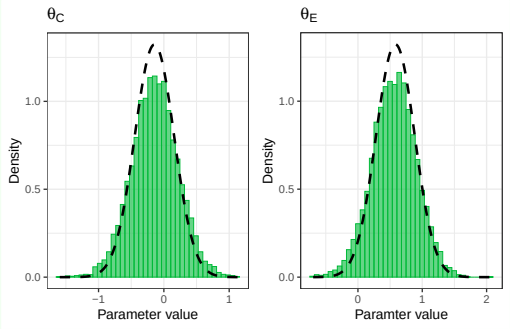
Example 4: Bayesian analysis using Markov Chain Monte Carlo computation

In Example 3 the posterior distributions for the mean parameters were calculated by analytical computation using the true SDs of the data. If, instead of the true SDs, the empirical SDs from the data $\hat{\sigma}_C \approx 1.075$ and $\hat{\sigma}_C \approx 1.200$ were used, this would lead to slightly different posterior distributions of $\mathcal{N}(-0.143, 0.104)$ in the control group and $\mathcal{N}(0.560, 0.126)$ in the experimental group (instead of $\mathcal{N}(-0.145, 0.0.091)$ and $\mathcal{N}(0.582, 0.091)$). The difference is due to the overestimation of the true SDs in both groups, leading to a higher SD and more weight of the prior distribution (larger ESS) in the posterior parameters as compared to the case with the true SD. In practice, the true SD is unknown, and setting the SD to a fixed value ignores the researcher's uncertainty about the parameter, as illustrated in Example 1.

As an alternative, a completely Bayesian model can be fit, where both the mean and the SD are modeled as random variables. In a completely Bayesian model, the posterior distributions are determined by both the prior distribution for the mean and the prior distribution for the SD. Depending on the choice of prior distributions, the posterior distributions in a completely Bayesian analysis can again be calculated by analytical methods using Bayes rule. For example, with the choice of uniform distributions for θ_i and $\log \sigma_i$, $i = C, E$, the marginal posterior distribution of the mean (averaged over the SD) is a t-distribution from the location-scale family of t-distributions and the marginal posterior distribution of the SD (averaged over the mean) is a scaled inverse-χ^2 distribution [59, Chapter 3.2]. What makes the application of completely Bayesian analysis difficult is that, with different choices of prior distributions or with additional model parameters, also the posterior distributions change and may become quite complex. Hence, in the context of designing and analyzing translational animal experiments, a more flexible framework than analytical computations is desirable, that does not require profound statistical knowledge and by-hand determination of the posterior distributions. Such a framework is provided by Markov Chain Monte Carlo (MCMC) computation. Fitting the model (1) in a completely Bayesian framework with MCMC computation leads to non-parametric representations of the posterior distributions for θ_i and σ_i, $i = C, E$, as MCMC draws presented in Figure 4.



(a) Partly Bayesian model where the SDs are modeled as fixed $\sigma_i = 1$, $i = C, E$.



(b) Comparison with a completely Bayesian model with random $\sigma_i \sim \mathcal{HN}(1)$ (half-normal distribution with scale parameter 1).

Figure 4: Posterior distributions of the group means θ_C and θ_E in a hypothetical new experiment with prior distributions $\theta_i \sim \mathcal{N}(0, 1)$. Green: approximate distribution as histogram of Markov Chain Monte Carlo (MCMC) draws. Dashed curve: exact distribution in a simpler analytic calculation with fixed SDs.

The histograms in Figure 4a represent the relative frequencies of the posterior draws of four MCMC chains, with 8000 post-warm-up iterations each and with the same prior distributions for the mean parameters $\theta_i \sim \mathcal{N}(0, 1)$, $i = C, E$, as in Example 3. In Figure 4b the partly Bayesian design with fixed SD is compared to the posterior distribution in a completely Bayesian analysis, where the residual standard deviations (SDs) are not fixed but are modeled as random variables from a half-normal distribution with scale parameter 1, as $\sigma_i \sim \mathcal{HN}(1)$, $i = C, E$.

Compared to the computation in a partly Bayesian design with fixed SD, the computation in a completely Bayesian design leads to larger tails of the posterior distributions of the group means. This is because, in the model with random SDs, the posterior distributions of the means are derived as marginal distributions by averaging over the posterior draws for the SDs. In contrast, in the model with fixed SDs, the posterior distributions of the means are only derived at one specific fixed value. Since point estimates in translational animal experiments are often unreliable, due to the small sample sizes, the model with fixed SDs may lead to overly confident results that don't reflect the researcher's actual uncertainty and completely Bayesian analyses may be preferred.

Prior distributions for the historical data

Variance parameters

With few data, special care has to be paid to the specification of a prior distribution for the heterogeneity parameter τ in the meta-analysis model (9) [37, 91]. It is recommended to use a prior distribution that puts most of its probability mass to areas that represent small to large heterogeneity and leaves only a small fraction to values that represent a larger heterogeneity [37]. The notion of "small" and "large" heterogeneity varies with the scale of the outcome of the animal experiment [37]. The classification of heterogeneity in context with the residual standard deviation σ of the data can provide a rough guideline for different heterogeneity degrees [98], as presented in Table 5.

Table 5: Classification of heterogeneity in relation to the standard deviation of the data, based on Weber et al. [98, p.28].

Heterogeneity	$\frac{\tau}{\sigma}$
small	0.0625
moderate	0.125
substantial	0.25
large	0.5
very large	1.0

Recommended weakly informative prior distributions for the heterogeneity parameter τ are then from the family of folded non-central t-distributions with special cases of the half-t and half-normal distribution, which are appropriate also for the challenging situation of few (below five) historical experiments or groups [37, 91]. In contrast, an uninformative uniform distribution on the heterogeneity parameter tends to lead to implausible high estimates in the situation of few experiments, and a conditionally conjugate inverse-Gamma prior on τ^2 is very sensitive to the choice of the distribution parameter in the situation, where the heterogeneity is estimated to be near zero [91]. Consequently, to be suited also for the situation of few historical experiments, a prior for τ is chosen from the family of non-central t-distributions in this thesis. The probability distribution functions of several candidates are plotted and their distributions across the heterogeneity classes from Table 5 are examined. For the heterogeneity classification, an estimate of the empirical SD of the historical data is used. The scale parameter is then chosen with the aim that the corresponding distribution of τ puts most of its probability mass (about 90%) to areas that represent small to large heterogeneity and leaves only a small fraction to values that represent a larger heterogeneity [37].

Mean parameters

For the mean parameter, the family of normal distributions is considered, which is the conditionally conjugate family to the normal likelihood distribution of the data [59, Chapter 3]. As a starting point for the situation of no other a priori information, a UIP in an empirical Bayes approach is considered [99, 100]. The mean of the UIP in the meta-analysis model (9) is set to an estimate of the mean in the historical data, which is called "empirical" Bayes since the distribution of the analysis data themselves is used to estimate parameters of the prior [100]. Alternatively, the prior parameters may be set according to estimates from related data. For example, estimates from wild type animals as negative controls could be used as starting point for the definition of a prior distribution for positive controls, representing a pathological condition, or estimates from younger animals could be used as prior distributions for older control animals. In an empirical Bayes approach, the empirical means $\overline{y}_g$ and variances $\hat{\sigma}_g^2$ (as in (4) and (5) but with group indices $g = 1, \ldots, G$ instead of only the two treatment groups $i = C, E$) may be used as reference for the prior definition. These estimates of the individual experiments can be combined to common meta-analysis (MA) estimates by weighting the individual experiments [101]:

$$\hat{\mu}_{\mathsf{MA}} = \frac{\sum_{g=1}^{G} w_g \overline{y}_g}{\sum_{g=1}^{G} w_g} \tag{14}$$

$$\hat{\sigma}_{\mathsf{MA}} = \frac{1}{N-G} \sum_{g=1}^{G} (n_g - 1)\hat{\sigma}_g^2 \quad g = 1, \ldots, G \tag{15}$$

The weights w_g in (14) and (15) may, for example, be chosen as inverse-variance weights $w_g = 1/\sigma_g^2$, $g = 1, \ldots, G$, to give the more precise estimates a larger weight. As an alternative empirical Bayes approach, the parameters for the prior distribution may be derived from the marginal distribution of the data under the assumption that the experiment-specific SDs are equal [99].

If the historical data are available on an individual animal level, the prior distributions for the historical data model may be based on estimates from a frequentist regression model, under the consideration of possible heterogeneity. An advantage of the MA estimates (14) and (15) derived from aggregated data is, that the historical has to be available only on a summary instead of on an individual animal level. Such summary estimates may be easier to share and store. On the other hand, the individual animal-level data is most informative and can reveal effects in specific subgroups. The differences between aggregate data meta-analysis (ADMA) with summary data and individual patient data meta-analysis (IPDMA) with individual animal data are further discussed in Section 4.1.1.

Reviewing the prior distributions of the parameters may be a first step to judge their plausibility based on a researcher's expectations of how typical data in this context should look like. However, these expectations are typically easier to formulate on the level of hypothetical animal data instead of on the level of the model parameters [102]. This is the idea behind prior predictive checks [59, Chapter 6.4],[103]. More specifically, in prior predictive checks, hypothetical data are simulated from the specified prior distributions, and the plausibility of the prior distributions is judged through the generated data.

The aim is then to decide if the model and parameter distributions seem plausible from a biological or medical perspective, based on the simulated data [1]. The theory is that, when the simulation is repeated many times for a large number of generated data sets, their empirical distribution approximates the true prior predictive distribution [103]. In practice, then, whether the simulated replication data from prior predictive checks seem reasonable or not has to be judged in light of the biological context. For example, one might have a rough guess about what outcomes are plausible for the force in a grip strength test or the protein level or protein-positive area in blood or histological samples.

Example 5: Prior distributions for a Bayesian model of the historical data

In Example 2 a normal-normal hierarchical model (NNHM) was fit to hypothetical historical control data. To fit the model in a Bayesian framework, prior distributions had to be specified for the overall mean μ, the heterogeneity parameter τ, and the residual SDs σ_g, $g = 1, \ldots, G$. It was assumed that the residual standard deviations σ_g were all equal to the same value σ. No prior distributions had to be specified for the experiment-specific means, θ_g, since their distributions were determined through the models for μ and τ via the relation in equation (9). The prior distributions for the remaining parameters were chosen as weakly informative empirical prior distributions with parameters as presented in Table 6. The resulting prior distributions are shown in Figure 5.

Table 6: Candidates for empirical prior distributions in a Bayesian meta-analysis (MA) of the historical data as normal distributions $\mathcal{N}(m, s^2)$ with mean m and variance s^2 and half-normal distributions $\mathcal{HN}(s^2)$ with scale parameter s and frequentist MA estimates $m_\mu = \hat{\mu}_{\text{MA}}$, $s_\mu = \hat{\sigma}_{\text{MA}}$ as in (14), (15).

Variable	Distribution	Choice of parameters
μ	$\mathcal{N}(m_\mu, s_\mu^2)$	$m_\mu = \hat{\mu}_{\text{MA}}$, $s_\mu = \hat{\sigma}_{\text{MA}}$
τ	$\mathcal{HN}(s_\tau^2)$	$s_\tau = \hat{\sigma}_{\text{MA}}/2$
σ	$\mathcal{HN}(s_\sigma^2)$	$s_\sigma = \hat{\sigma}_{\text{MA}}$

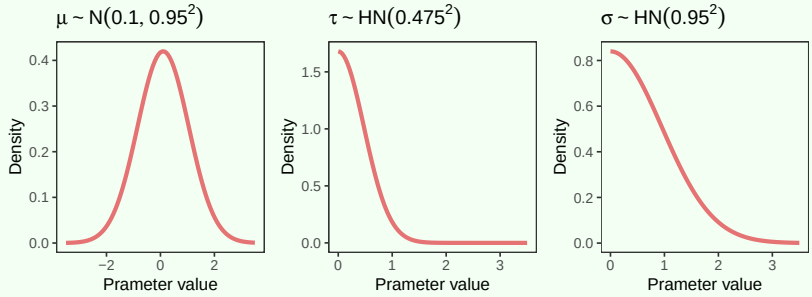


Figure 5: Prior distributions in the meta-analysis model of the historical data. Shown are the probability density functions for the overall mean μ, residual SD σ, and the heterogeneity τ of the experiment-specific means.

The scale parameter for τ was set, so the corresponding distribution puts most of its probability mass (83%) to heterogeneity classes in the range of small to large heterogeneity (where "small" and "large" refer to the classification in Table 5 with $\hat{\sigma}_{\text{MA}}$ as reference scale) and leaves only the remaining 17% to larger heterogeneity classes.

Example 6: Prior predictive checks for the Bayesian model of the historical data

To judge the plausibility of the prior distributions from Example 5, graphical prior predictive checks are performed. To this end, twelve replicate data sets are simulated from the specified prior distributions of μ, τ, and σ via the NNHM in (9), as shown in Figure 6.

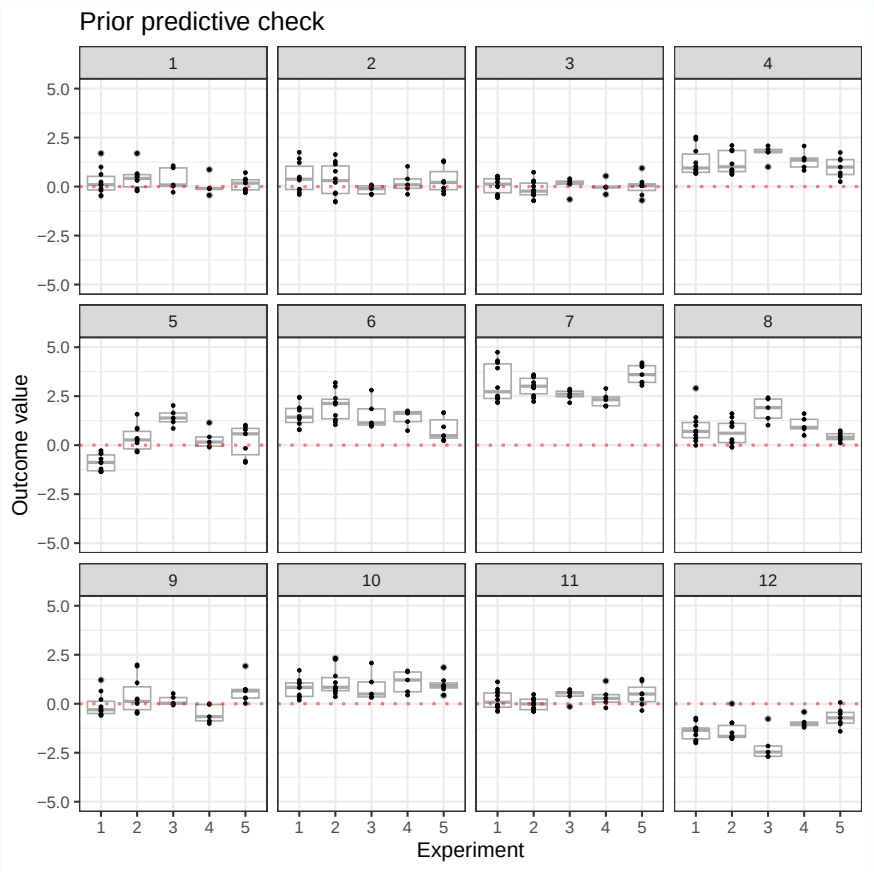


Figure 6: Prior predictive check, as twelve hypothetical data sets simulated from the candidate prior distributions in the Bayesian model of the historical data. Each data set consists of five experiments (box plots) with individual animal data (black dots) and sample sizes as in the historical data. The overall mean is shown as a dotted red line.

In comparison with the observed historical data from Example 2, the chosen prior distributions seem to produce reasonably similar data. The spread of the replicated data is wider than in the simulated data, which is intentional since the prior distributions for the historical data are meant to represent only a rough guess of plausible ranges but not the data themselves.

Predictive distributions for the new data

In Section 2.2.1 it has been proposed to fit a meta-analysis model to available historical data. The resulting estimates of the mean and residual standard deviation are useful to make better assumptions for the design and sample size planning of a new animal experiment. In addition, they can also be used to formulate prior distributions for a Bayesian analysis of the new data [39, 42–45] with the aim to support the newly observed data by additional a priori knowledge and potentially reduce the necessary sample size. When a Bayesian model is fit to the historical data, estimates of the mean and variance parameters are given as posterior probability distributions. These posterior distributions are used to make predictions for the mean and standard deviation of the control group in the new experiment, as explained in the following section and illustrated in Figure 7.

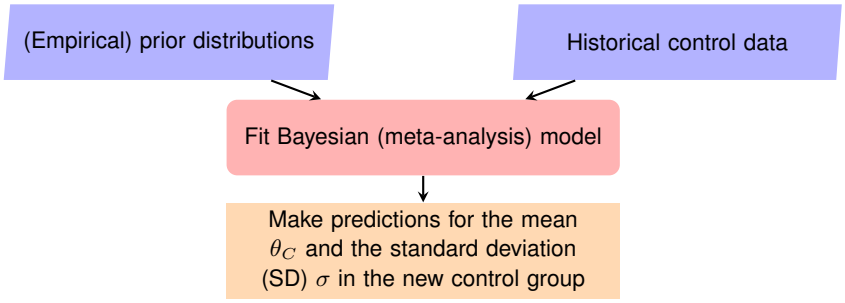


Figure 7: Prediction part of the sample size planning procedure with Bayesian predictive approaches.

Posterior predictive distribution of new animals

When planning a new experiment, the researcher is interested in making statements about the measurements in the future experiment as the basis for a sample size decision. The distribution of a measurement y^* in an animal in the new experiment, given the data of historical animals, $y_1, \ldots, y_n$, is referred to as *posterior predictive distribution of a new observation* and denoted as $p(y^*|y_1, \ldots, y_n)$. This distribution depends on the assumed relationship between the historical and the new data, as illustrated in Table 3.

For a normally distributed historical data set, consisting of independent and identically distributed measurements $y_k \sim \mathcal{N}(\theta, \sigma^2)$, $k = 1, \ldots, n$ with sample mean $\overline{y}$ and a fixed σ, and with a normal prior distribution for the mean parameter $\theta \sim \mathcal{N}(m, s^2)$ sample size n and a fixed σ, the posterior predictive distribution for a potential measurement y^* in the new experiment is (under

the assumption of exchangeability (similarity) of the historical and the new measurements) again a normal distribution with parameters [59, Chapter 2]

$$\begin{aligned}\mathsf{E}(y^*|y_1,\ldots,y_n) &= m^*\\ \mathsf{Var}(y^*|y_1,\ldots,y_n) &= \sigma^2 + (s^*)^2,\end{aligned}$$

where m^* and s^* are given as in (12) and (13). If both the mean and the SD parameter are unknown and modeled as random variables, instead of fixed values, or if the model includes further parameters, the probability of different future measurements can still be calculated by integrating over the unknown model parameters. When the calculations become more complex, draws of the posterior predictive distribution can be obtained by simulation from MCMC draws of the joint posterior parameter distribution, as discussed in Section 2.2.3.

The posterior predictive distribution of a control animal in the new experiment y_C^* can be used to judge the plausibility of the posterior parameters in prior predictive checks (as explained in the previous section). Moreover, it could be used in the context of sample size estimation to simulate hypothetical data for a new experiment and calculate the power of Bayesian hypothesis tests as the average of rejected null hypotheses over a large number of simulated data sets. The experimental group could be simulated from a normal distribution with a mean that is obtained from the mean m^* by adding a relevant mean difference δ and variance σ^2 or as sum of the σ^2 plus the variance in the posterior estimation of the model parameters. A similar procedure will be suggested in Section 2.3. Instead of modeling both the residual variance and the variance in the posterior estimation of the model parameters, however, the data are simulated in a frequentist approach from a normal distribution that considers only the residual variance. This allows better control of the experimental design and better reflects the hypotheses, that are formulated with fixed parameters (which is necessary to guarantee a correct interpretation of the Bayes factors [104]), as further discussed in Section 4.1.2.

While the uncertainty in the posterior distributions is not considered in the simulation of the hypothetical new animal data, it is incorporated into the definition of prior distributions for the mean parameter θ_C^* of the new control group. The *posterior predictive distribution for the mean parameter*, given the historical data, $y_1,\ldots,y_n$, is denoted as $p(\theta_C^*|y_1,\ldots,y_n)$ and represents the expected value of the posterior predictive distribution of new observations. The residual variance σ^2, which was important in the context of posterior predictions on the level of the observations in the previous paragraph, drops out in the expected value. Depending on the assumptions of the relationship between the historical data and new control data, Table 3 is extended by different prediction models as presented in Table 7.

Table 7: Models for the predictive distribution of the mean θ_C^* in a new experiment, that are computed from the posterior MCMC draws of the historical data model (based on [32, Section 3.16, p.90f] and [39, 77]). μ_r, τ_r, θ_r and β_r, $r = 1, \ldots, R$, denote the posterior MCMC draws of the common mean, the heterogeneity, the group means and further regression coefficients in the Bayesian model of the historical data and θ_{Cr}^* the generated draws for the future mean. R is the total number of the MCMC draws after warm-up.

Model	Assumption on the new mean w.r.t. the historical means	Model from the posterior draws of the historical data
(a)	Identical as all the other means in the historical data	$\theta_{Cr}^* = \mu_r$
(b)	Differs from the overall historical mean μ by the (fixed) population effect β of some covariate X	$\theta_{Cr}^* = \mu_r + \beta_r x^*$, where x^* is the value of X in the new experiment
(c)	Identical as the mean θ_g of a specific, predetermined historical subgroup $g \in \{1, \ldots, G\}$, where the G subgroup means vary randomly around the overall mean	$\theta_{Cr}^* = \theta_{gr}$, where g is fixed for all r
(d)	Represents the variation across all the historical means from (c)	$\theta_{Cr}^* = \theta_{g(r)}$, where $g(r) \in \{1, \ldots, G\}$ is random and may vary with r
(e)	Brand new but related to the historical subgroup means from (c) by the assumption $\theta_1, \ldots, \theta_G, \theta_C^* \sim \mathcal{N}(\mu, \tau^2)$	$\theta_{Cr}^* \sim \mathcal{N}(\mu_r, \tau_r^2)$
(f)	Independent of the historical means	no model

As mentioned before, the focus of this thesis lies in the models (a), (b), (e) and (f). Predictions in the models (a) - (e) differ with respect to their variability (uncertainty) and in model (f) no predictions are made. In addition, in model (b) also the expected value differs from the means from the other models. In model (a), the uncertainty in the prior distribution corresponds to the uncertainty in the posterior distribution of the overall mean, while in model (b) the uncertainty in the prediction corresponds to a sum of both the uncertainty in the posterior distribution of the overall mean *and* the covariate effect, leading to a wider posterior predictive distribution. Also model (e) has a larger uncertainty in the predictions than model (a) due to both the sampling uncertainty, which is represented by modeling the new mean as samples from a normal distribution with between-group variance τ^2, and the uncertainty in the posterior estimation of μ and τ. Similarly, the variance in the posterior predictive distribution of a new observation is a sum of the residual variance σ^2 and the variance s^* of the posterior distribution of the mean (that is due to the limited sample size of the historical data). With an increasing sample size, the variance in the posterior distribution of the mean decreases, while the residual variance σ^2 remains [59, p.357].

In addition to these models, an "equal-but-discounted" approach [32, Section 3.16] can be realized by taking the predictive prior from (a) and scaling it by the likelihood distribution of the historical data raised to some "discounting" parameter smaller than one, which has the effect of decreasing the weight of the historical data in the analysis of the new data [8]. Moreover,

a "biased" [32, Section 3.16] approach could be realized by modeling the historical means as a biased version of the mean in the new experiment, where the bias is set in accordance with the experimental researcher. Disadvantages of these approaches are that they require an explicit specification of the discounting or bias parameters, whereas in the models from Table 7 the parameters are estimated from the data. Alternative prediction models are further discussed in Section 4.1.1.

For the residual SD, the analogous assumptions can be made by replacing θ with σ in Table 7.

Example 7: Predictive distribution for the new control animals

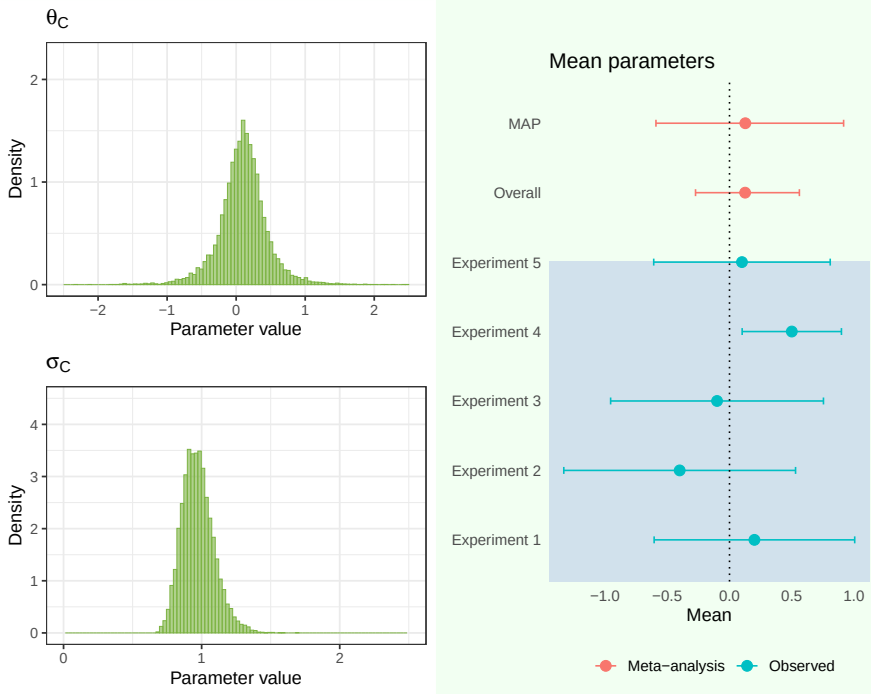
In Example 2 a Bayesian hierarchical (meta-analysis) model was fitted to the historical experiments. The model provides posterior distributions for the overall mean μ, of the between-experiment variability (heterogeneity) τ, of the experiment-specific-means θ_g, $g = 1, \ldots, 5$ and of the common SD σ. These posterior distributions can be used to make predictions for the mean and SD parameter in the new experiment. The first five MCMC draws of the posterior distributions of μ, τ and σ are presented in Table 8.

Table 8: First five MCMC draws $r = 1, \ldots, 5$ of the posterior distribution of the overall mean μ, SD σ and heterogeneity τ of the means in a Bayesian hierarchical model of the historical control data. The columns θ_C and σ_C on the right are the predictions for the mean and SD in the new control group assuming similarity for the means as $\theta^*_{Cr} \sim \mathcal{N}(\mu_r, \tau_r^2)$, and equality for the SDs as $\sigma^*_{Cr} = \sigma_r$, $r = 1, \ldots, 5$.

Draw	μ	τ	σ	θ^*_C	σ^*_C
1	0.32	0.17	0.87	0.46	0.87
2	0.04	0.25	0.98	0.20	0.98
3	0.23	0.32	0.98	0.51	0.98
4	0.05	0.11	1.06	0.12	1.06
5	0.06	0.14	0.85	0.24	0.85

The predictive distribution for the mean θ_C for the new experiment is calculated from the posterior MCMC draws of the overall mean μ and the heterogeneity τ by drawing for each MCMC draw $r = 1, \ldots, 3000$ one sample from the distribution $\mathcal{N}(\mu_r, \tau_r^2)$ (where μ_r is the r'th MCMC draw of the overall mean and τ_r is the r'th MCMC draw of the heterogeneity). For example, in the first MCMC iteration, the posterior draws are $\mu_1 = 0.32$ and $\tau_1 = 0.17$. Under the model (e) from Table 7, a random value for the new mean is then sampled from the distribution $\mathcal{N}(0.32, 0.17^2)$. The (random) result is here $\theta^*_{C1} = 0.46$ (as first entry in the column θ^*_C in Table 8). In the second MCMC iteration, the posterior draws are $\mu_2 = 0.04$ and $\tau_2 = 0.25$, so the next value for the new mean is sampled from the

distribution $\mathcal{N}(0.04, 0.25^2)$. The (random) result is $\theta_{C2} = 0.20$. The procedure is repeated for all other MCMC iterations, leading to 3000 draws for θ_C^*. The predictive distribution for θ_C^* then consists of all these draws, as shown in the upper part of Figure 8a. The SD σ_C^* of the control group of the new experiment is predicted under a model corresponding to the assumption (a) in Table 7. Hence, the prediction is identical to the posterior distribution of the historical SD σ, and the columns σ and σ_C^* in Table 8 are the same. Alternatively, deviating SDs of the new and historical data may be chosen according to some other model in Table 7. The predictive distribution for σ_C^* is shown in the lower part of Figure 8.



(a) MCMC draws (b) Forest-plot

Figure 8: Predictive distribution of the mean and standard deviation (SD) for a new experiment. (a) Histograms of the Markov Chain Monte Carlo draws representing the predictive distribution for the new control group's mean θ_C and SD σ_C. (b) Forest-plot of the median and 95% equal-tailed interval (ETI) (as quantiles of the posterior distribution) of the predicted control mean together with the median and ETI of the overall mean (red) and frequentist experiment-specific mean estimates from a stratified analysis of the individual experiments, as arithmetic means and frequentist 95% confidence intervals (blue).

While the predictive distribution for the SD is modeled to be identical to the posterior distribution from the common historical SD, the predictive distribution of the mean has a

wider variance than the posterior distribution of the overall mean, as illustrated in Figure 8b. This is because the overall mean only reflects the central tendency of the experiment-specific means, whereas the MAP also reflects the fact that an experiment-specific mean may deviate from the overall mean by a variability according to τ.

Approximation by parametric distributions

The predictions for the new experiment, which are calculated from the posterior MCMC draws, are available as series of MCMC draws. To use the predictions as prior distributions in the analysis of the new experiment, a more convenient representation than the non-parametric distribution of MCMC draws is desirable. Therefore, the series of MCMC draws are approximated by parametric distributions [39]. The normal distribution family is chosen as a parametric family for the mean parameter in the new experiment since it is a relatively simple and well-known distribution. Formally, in case of the NNHM from equation (9) and with a normal prior distribution for μ and a half-normal prior distribution for τ, the predictive distribution for the mean parameter μ and the parameter θ^* in the new experiment are normal-mixture distributions [105]. Thus, if a very close fit to the posterior distribution is desired, more complex distribution families have to be considered for the approximation of the posterior MCMC draws, as discussed in Section 4.1.1. After a parametric distribution family is specified, its parameters can be determined, e.g., by maximum likelihood (ML) estimation or by the expectation-maximization (EM) algorithm [106].

Effective sample size of the prior

Equation (13) showed that, conditionally on a fixed residual standard deviation σ and using normal prior distributions for the mean θ as conditionally conjugate prior distributions, the posterior precision $1/(s^*)^2$ is equal to the sum of the prior precision $1/s^2$ (where s is the SD of the prior for θ) plus the precision of the data $1/(\text{SE}(\overline{y})^2$ as square of one divided by the standard error of the mean ($\text{SE}(\overline{y}) = \frac{\sigma}{\sqrt{n}}$). To get an idea of how many observations the prior distribution is equivalent to, the prior precision can be defined in terms of the the variance of data as $1/s^2 = n_0/\sigma^2$ to have the same denominator as the data precision. Then, the term $n_0 = \sigma^2/s^2$ is referred to as *effective sample size (ESS) of the prior distribution* for the mean parameter. In the context of planning and analyzing a new animal experiment with historical control this implies that, for a fixed residual SD, the prior distribution (that was predicted from the historical data model) with SD s_C can be said to be equivalent to $n_{0C} = \sigma_C^2/s_C^2$ animals in the new experiment. When the variance parameter is modeled as a probability distribution, (13) no longer holds. Instead, the posterior distribution has

to be calculated by considering the joint distribution of the mean and residual. In the context of a meta-analysis, additional variation in terms of the heterogeneity of subgroups comes into play. There, an equivalent effective sample size of the prior can be computed as the ratio of the prior distribution's variance in a heterogeneous sample of historical data over a homogeneous, pooled sample.

More generally, in this scenario, the effective sample size of the prior $p(\theta_C)$ can be calculated as a predictively consistent effective sample size, where the predictive consistency criterion ensures that the expected posterior ESS is equal to the sum of the prior ESS and the sample size of the observed data [107]. The number n_{0C} provides as *maximum prior effective sample size* an upper bound for the predictively consistent ESS for the case of infinite historical information (in terms of the sample sizes and the number of historical experiments approximating infinity, [39, 107]). For the predictively consistent ESS approaching the value one, the collection of additional data does not improve the inference with the corresponding prior as compared to the ideal case of infinite information. The maximum prior effective sample size is used to categorize the distribution of the heterogeneity parameter in Table 5.

Example 8: Parametric approximation of the predictions

In Example 7, predictive distributions were constructed from the posterior distributions in the model of the historical control data presented in Example 2. This led to a non-parametric representation of the predictive distributions as series of MCMC draws. To transform these series of draws into a format, that can straight-forward be set as prior distributions for the mean and SD parameter in the new experiment's control data, the draws are approximated by parametric distributions.

According to theory, the predictive distribution for the mean parameter from the NNHM is, with a normal prior for the overall mean μ and a half-normal prior for the heterogeneity τ in the historical data, as in Example 6, a normal mixture distribution [105]. Here, to keep the interpretation and handling of the prior distributions simple and to avoid overfitting to the historical data, simple normal distributions (as special case of normal mixture distributions with only one single component) are used for the approximation of the predictive distribution of the mean and the logarithmic SD. The predictive distribution of the SD is approximated by a log-normal distribution because it is skewed and has only positive values. The MCMC distributions are shown in Figure 9 together with the fitted parametric distributions. If a closer fit to the historical data is desired, mixture distributions may be used, as shown by the dashed curve in Figure 9 for the mean parameter (with parameters fitted via expectation-maximization (EM) and the number of mixture components chosen

to minimize the Akaike information criterion (AIC)). In contrast to the prediction for the SD, in the prediction model for the mean, the heterogeneity in the group means is reflected. This results in a more complicated predictive distribution and a closer fit of the mixture distribution than of the simple normal distribution.

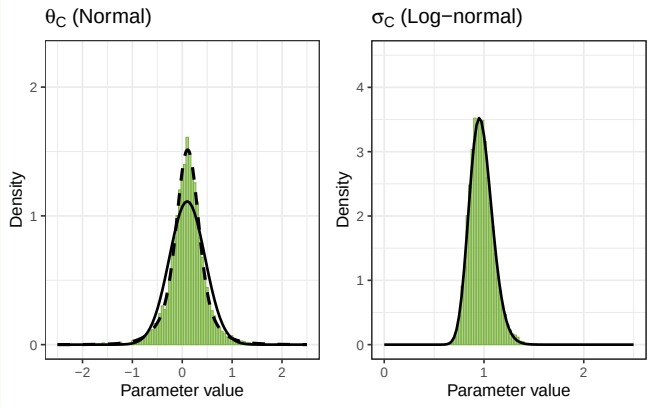


Figure 9: Histogram of MCMC draws (green) and curves of the fitted distributions normal and log-normal distributions (black solid lines) of the predicted mean (left) and SD (right) of the new control data. For comparison, the fit of a normal mixture distribution (with the number of components chosen to minimize the Akaike information criterion (AIC) in a sequence of candidates from one component to four components) is shown for θ_C as indicated by the dashed lines.

Summary 1: Predictive prior distributions

In this thesis, the derivation of prior distributions for a new animal experiment from historical data involves the following steps:

(1) Fitting a Bayesian model to related historical data (using Markov Chain Monte Carlo (MCMC) simulation)

(2) Choosing a prediction model for the new experiment (based on the assumed relationship between the historical data and the data in the new experiment)

(3) Using the estimated posterior distributions (which are here calculated as series of MCMC draws) from the model fitted in (1) to make predictions for the new data according to the model assumed in (2)

(4) Fitting a parametric distribution to the series of MCMC draws of the predictive distributions from (3)

Random variation in the experiments' parameters (heterogeneity) may, e.g., occur in cases of different experimental laboratories or batches in the historical experiments. In this case, the predictive distribution for the parameters in the new experiment (as mean and SD for normally distributed data) will typically be wider than the posterior distribution modeling the average of the experiment-specific parameters. If only data from a single historical experiment or group is available, no heterogeneity can be estimated. Under strictly controlled conditions, predictions may be made under the assumption that the mean and SD in the historical and the new data coincide.

2.3 Simulation based design analysis and sample size determination

The following section refers to the second part of the sample size planning procedure with Bayesian predictive approaches and is subdivided into two parts. In the first part, methods for testing Bayesian hypothesis are discussed. The choice of hypothesis test is important when making sample size decisions with regard to the statistical power and type I error rate of the test. In the second part of this section, it is explained how simulations can be used to examine varying experimental designs and to decide on a final sample size for the new experiment. The presented workflow for the design simulation and sample size determination represents a formalization of my previous contribution [1] and may be used for different assumptions on the relatedness of the historical and the new data.

2.3.1 Concepts of Bayesian hypothesis testing

In the frequentist framework, that was described in Section 2.1.1, the sample size was oriented at a statistical hypothesis test for the null hypothesis. There is a standard procedure for the construction of frequentist tests for a null hypothesis H_0. More specifically, frequentist tests are commonly constructed as *uniformly most powerful tests* [71, Chapter 3], so that they have maximal power among all tests for H_0 at a certain significance level α. This leads to the two-sample t-test (3) in case of two normally distributed groups with equal variances [71, Section 5.3]. In a Bayesian framework, there is no standard approach for the construction of a hypothesis test [53]. In this thesis, the focus lies on hypothesis tests that are based on Bayes factors, which provide several conceptual advantages over frequentist testing and other Bayesian decision rules [51, 54] and are suggested as default choices for Bayesian hypothesis testing by some authors [108,

109]. Additionally, hypothesis tests, that are based on the credible intervals of the posterior probability of the mean difference being different from zero, are investigated [8]. The following section refers to the hypothesis testing part of the sample size planning procedure and is illustrated in Figure 10.

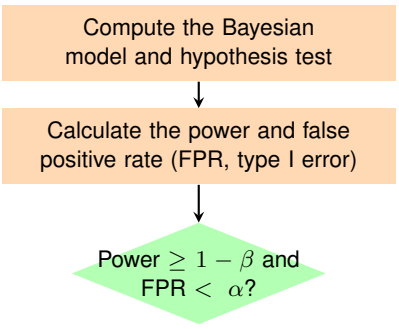


Figure 10: Hypothesis testing part of the sample size planning procedure with Bayesian predictive approaches.

Testing hypothesis based on credible intervals

In frequentist analysis, confidence intervals are used to quantify the uncertainty in the estimation of a parameter. Moreover, confidence intervals can be used as equivalents to hypothesis tests for a null hypothesis at a significance level α by rejecting the null hypothesis if the null value is not contained in the $1-\alpha$ confidence interval for that parameter [71]. For example, $H_0 : \delta = \delta_0$ is rejected if δ_0 is not contained in the $1-\alpha$ confidence interval for δ.

Credible intervals are Bayesian versions of frequentist confidence intervals. They have slightly different interpretations than frequentist confidence intervals. By definition, a frequentist $1-\alpha$ confidence interval for a parameter δ is a random interval $\text{CI}_{1-\alpha}^{\text{Frequ}} = (L_\alpha(Y), U_\alpha(Y))$ with random bounds, that covers the unknown (but regarded as fixed) parameter δ with probability $1-\alpha$:

$$P\left(\delta \in (L_\alpha(Y), U_\alpha(Y))|\delta\right) = 1-\alpha. \tag{16}$$

Hence, the idea of a confidence interval is, that if a researcher would examine many data sets from the same probability model (that depends on (the fixed) parameter δ) and if the researcher would calculate a confidence interval for each data set, then, on average, $(1-\alpha)\%$ of these intervals would cover the true parameter δ. The interpretation of a $1-\alpha$ Bayesian credible interval $\text{CI}_{1-\alpha}^{\text{Bayes}} = (l_\alpha, u_\alpha)$ is, instead, that it is set up with bounds $l_\alpha = l_\alpha(y)$ and $u_\alpha = u_\alpha(y)$ that

are regarded as fixed so that the posterior probability (given the data) of a random realization of δ falling within the credible interval is $1-\alpha$. This is (with $p(\delta|Y=y)$ denoting the value of the posterior probability density function of δ, given the data y) formally expressed as [110, Chapter 6]

$$\int_{l_\alpha}^{u_\alpha} p(\delta|Y=y)\,\mathrm{d}\delta = 1-\alpha. \tag{17}$$

This definition implies that the Bayesian $1-\alpha$ credible interval $I_{1-\alpha}$ includes $(1-\alpha)\%$ of the probability mass of δ, given the data. This definition does not uniquely define the credible interval, so there are multiple options to construct a credible interval in practice. A common type of Bayesian credible interval is the quantile interval or equal-tailed interval (ETI). The bounds of an $1-\alpha$ posterior ETI for a parameter δ are given by the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of its posterior probability distribution

$$\mathsf{CI}_{1-\alpha}^{\mathsf{ETI}} = \left(q_{\delta|y}\left(\frac{\alpha}{2}\right), q_{\delta|y}\left(1-\frac{\alpha}{2}\right)\right), \tag{18}$$

where $q_{\delta|y}(z)$ denotes the z-quantile, $z\in[0,1]$, of the posterior probability distribution of δ, given the data. When the posterior distribution is calculated with MCMC methods, then the quantiles are computed as sample quantiles of the distribution of MCMC draws as

$$\hat{q}_{\delta|y}(z) = (1-\gamma)\delta_{[j]} + \gamma\delta_{[j+1]}, \tag{19}$$

where $\delta_{[1]} \leq \delta_{[2]} \leq \ldots \delta_{[R]}$ are the R ordered MCMC draws of the posterior distribution of δ (after warm-up of the MCMC chains), that correspond to the R fractions $z_r = (r-1)/(R-1)$, $r=1,\ldots,R$, of the samples from the posterior distribution of δ for the draws. By definition, quantiles divide the posterior distribution into two parts: a lower part consisting of the $z\%$ of all posterior samples, that are smaller or equal than $\hat{q}_{\delta|y}(z)$, and an upper part consisting of the $(1-z)\%$ of all posterior samples, that are greater than $\hat{q}_{\delta|y}(z)$. Hence, the value j in (19) denotes the greatest integer less than or equal to $R\cdot z+(1-z)$ as $j=\lfloor R\cdot z+(1-z)\rfloor$. For values z that don't correspond directly to a fraction z_r, $r=1,\ldots,R$, the value $\gamma = R\cdot z+(1-z)-j$ denotes the fraction by which the value z lies on the way from z_r to z_{r+1}.

An alternative commonly used credible interval is the highest posterior density interval (HDI), which is defined as a credible interval for which the probability density (indicating how likely different parameter values are) of each parameter value lying inside the interval is higher than the probability of any parameter value lying outside the interval:

$$p(\delta|y) \geq p(\delta'|y) \quad \forall\, \delta \in \mathsf{CI}^{\mathsf{HDI}}_{1-\alpha}, \quad \delta' \notin I^{\mathsf{HDI}}_{1-\alpha}. \tag{20}$$

The differences between HDIs and ETIs regarding their practical use are discussed in Section 4.1.2. For either definition of the credible interval, a hypothesis test at level α can then be set up by checking if the Bayesian $1-\alpha$ credible interval excludes the null value δ_0 under H_0 (then the probability of δ being different from δ_0, given the data and prior distribution, is at least 95%) [8]:

$$D_{\mathsf{CI}} = \mathbb{1}\{\delta_0 \notin \mathsf{CI}^{\mathsf{Bayes}}_{1-\alpha}\}, \tag{21}$$

where $\mathbb{1}\{\delta_0 \notin \mathsf{CI}^{\mathsf{Bayes}}_{1-\alpha}\}$ denotes the indicator function, that takes the value one if the true value is included in the credible interval and zero if it is not included.

Example 9: Bayesian credible interval

In Example 3, data from a hypothetical new experiment with a control (C) and an experimental (E) group were simulated. The groups' means were assigned standard normal distribution prior distributions as $\theta_i \sim \mathcal{N}(0,1)$, $i = C, E$. Hence, the prior for $\delta = \theta_E - \theta_C$ was the difference of two normally distributed variables and therefore again normally distributed with mean $m_\delta = 0$ and variance $s^2_\delta = 2$. For simplification, the example is continued with the fixed residual SDs instead of modeling them as random variables in a completely Bayesian design. With residual SDs $\sigma_C = \sigma_E = 1$ the parameters of the posterior distribution were calculated in Example 3 as approximately $\mathcal{N}(-0.145, 0.0.091)$ for the control group and $\mathcal{N}(0.582, 0.091)$ for the experimental group. This results in a normal posterior distribution of δ as difference of the normal distributions of θ_E and θ_C as

$$\delta|y \sim \mathcal{N}(0.582 + 0.145, 0.091 + 0.091) = \mathcal{N}(0.727, 0.182).$$

For the level $\alpha = 0.05$, the posterior ETI for δ is then given by the quantiles of the distribution $\delta \sim \mathcal{N}(0, 2)$ as approximately

$$\text{CI}^{\text{ETI}}_{1-\alpha} = (-0.132, 1.532).$$

The HDI boundaries are calculated by optimization as those values that minimize the credible interval width among all 0.95% credible intervals for δ. In this case, the calculation results with $(-0.109, 1.563)$ approximately again in the same interval boundaries as for the ETI, which is in line with theory, since the HDIs and ETIs of symmetrical distribution should coincide. The prior and posterior intervals for all model parameters are plotted together with the data and prior and posterior distributions for θ_C, θ_E, and δ in Figure 11.

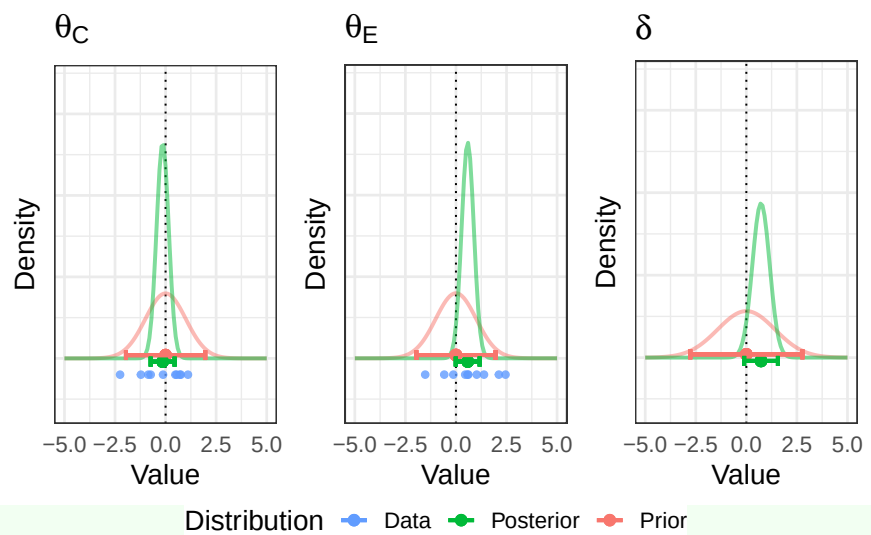


Figure 11: Prior and posterior density function for the distribution of the group means θ_C, θ_E and their difference δ together with the 95% highest posterior density intervals (HDIs) (horizontal lines) and the expected values (as dots) of the distributions.

Since the null value $\delta_0 = 0$ falls into the credible interval, $H_0 : \delta = 0$ cannot be rejected with the credible-based decision rule (21). The precision of the posterior mean is a sum of the precision of the prior and the data precision. Hence, a larger sample size will generally lead to a narrower posterior credible interval and, thus, to a higher "power" of this test.

Testing hypothesis based on Bayes Factors

In a Bayesian context, a general hypothesis test can be obtained by assigning the hypothesis pair in (2) prior probabilities and updating them by Bayes rule [111]. A problem with this general approach is, that the posterior probabilities of the hypothesis are sensitive to the choices of prior distribution for the hypothesis. As an alternative, it has been suggested to test hypotheses by the comparison of the probabilities of the observed data under two models, which here refer to the hypothesis pair from (2), using the Bayes factor

$$\mathsf{BF}_{10} = \frac{p(y|H_1)}{p(y|H_0)},$$

where y is the data in the new experiments (which are here the control and experimental data $y_{C1}, \ldots, y_{Cn_C}, y_{E1}, \ldots, y_{En_E}$). By definition, the Bayes factor is the ratio of marginal likelihood densities and indicates how much more likely the data are under H_1 than under H_0. The Bayes factor is thus interpreted as a relative measure of evidence for H_1 as compared to H_0. Each marginal likelihood is thought to indicate how reasonable a set of observed animal data appears under the two hypotheses. This allows statements of which hypothesis may correspond to the true distribution of the data. A value of the Bayes factor for the hypothesis (2) in an animal experiment that is greater than on indicates that, under consideration of the new experiment's data, there is more evidence for unequal group means (H_1) than for equal group means (H_0).

Here, the distribution parameters are the group means θ_C, θ_E, or rather the difference δ, and SDs σ_C, σ_E. The marginal likelihood densities $p(y|H_j)$, $j = 0, 1$, in the Bayes factor are calculated as weighted average of the conditional likelihood function of the data under the hypothesis H_j weighted by the prior distributions of the model parameters under the hypothesis H_j [111, 112]. Under $H_0 : \delta = 0$, the likelihood function is fixed at a specific value as $\delta = 0$ and the joint prior distribution of the model parameters reduces to a distribution of the SDs. In contrast, under the alternative $H_0 : \delta \neq 0$, the parameter δ may take any other value than zero, so the likelihood function is evaluated for an unknown δ and prior distributions are required for both the SDs and for δ [113].

The prior distribution for an arbitrary δ under H_1 is defined as average over the joint density of the model parameters under H_1:

$$p_1(\delta|H_1) = \int\int f_1(\delta, \sigma_C, \sigma_E)\mathsf{d}\sigma_C\mathsf{d}\sigma_E, \tag{22}$$

where $f_1(\delta, \sigma_C, \sigma_E)$ is the density function of the joint parameter distribution of the model parameters and is defined (as derivative of a mixture probability measure over H_0 and H_1, that is continuous, except from δ having concentrated prior probability at $\delta_0 = 0$ [114]) as in equation (2.14) by Dickey and Lientz [112]. Similarly, the posterior distribution $p_1(\delta|H_1, y)$ for an arbitrary δ under H_1 is defined as in equation (2.19) by Dickey and Lientz [112], taking the likelihood distribution of the data into account.

In the case of the point null-hypothesis $H_0 : \delta = 0$, and under the assumption that, as δ approximates the null value $\delta_0 = 0$ the prior distributions for the residual SDs σ_C and σ_E under H_0 are the same as the prior for the residual SDs under H_1 (which is fulfilled if the prior distributions for the mean and the residual SD are independent [115]), the computation of the Bayes factor can be simplified as Savage-Dickey density ratio [112]

$$\mathsf{BF}_{10} = \frac{p_1(\delta_0|H_1)}{p_1(\delta_0|y, H_1)}, \tag{23}$$

where $p_1(\delta_0|H_1)$ is the value of the prior distribution for δ under H_1 at the null value $\delta_0 = 0$ and $p_1(\delta_0|y, H_1)$ is the value of the posterior distribution of δ under H_1 at $\delta_0 = 0$ (defined as in the publication by Dickey and Lientz [112]). Equation (23) reduces the computation and interpretation of the Bayes factor to the ratio of the height of the prior probability density function of δ under H_1 at the null value $\delta_0 = 0$ over the posterior probability density function of δ under H_1 at the null value $\delta_0 = 0$. With this formulation, the value of the Bayes factor becomes large, if the consideration of the new animal data makes a group mean difference of zero appear much less likely than before the consideration of the data. Concerning the impact of different model components on the Bayes factor, the following statements can be made [115]:

1. The prior distribution for δ has a major impact on the Bayes factor.

    a) With the same prior but an increasing sample size, the impact of the prior on the posterior decreases. Thus, a change in the prior for δ has a larger impact for smaller sample sizes.

    b) With a flatter prior distribution for δ (that has larger distribution tails and is thus less informative), the height of probability density function at a specific null value δ_0 decreases. The reduced prior probability at δ_0 will typically entail also a reduced posterior probability at δ_0 as compared to the posterior value at δ_0 with a more informative prior distribution. However, in this situation, the change in the posterior at δ_0 will generally be smaller than in the change in the prior at δ_0, since the posterior is also

influenced by the data. Hence, when moving to less informative prior distributions, the nominator in the Bayes factor (23) decreases more than the denominator. As a result, the Bayes factor becomes smaller and tends to favor the null hypothesis, irrespective of the data, as described by Bartlett's paradox [116] (as a reformulation of the paradox that was originally described by Jeffreys and Lindley [55, 117]). Consequently, uninformative prior distributions can lead to test results in favor of H_0, irrespective of the data.

c) In the other extreme, however, if the prior distribution is too informative, the misspecification risk increases and may again lead to automatic preference of the null model [118].

2. The nuisance parameters (parameters of secondary interest, which here are the SD parameters σ_C and σ_E) affect the Bayes factor only implicitly (through the posterior distribution of δ, which is calculated as average of the joint parameter distribution over σ_C and σ_E).

These statements imply practical consequences in the context of the statistical planning and analyzing of translational animal experiments. Concerning item a) of the first statement, the fact that group sizes in translational animal experiments are typically small [7, 8], implies that the experiment's biological or medical conclusion is very sensitive to the choice of the prior distribution for the difference in treatment means. The difference in treatment means is mainly determined by the posterior distribution of the treatment means in the control and experimental groups (as noted by the second statement). These posterior means in turn are here determined by the information from the historical data. More specifically, predictions from available control data are used to formulate a prior distribution for the mean and SD of the control group in the new experiment (under the assumption of some relatedness of the historical and the new data). The prior distributions for the experimental group in the new experiment are then centered at the same values as the control prior distributions but are set up to have larger distribution tails, that give more weight to new evidence (with details explained in Section 2.2.3). In the model of the historical data, the prior distributions are chosen to let the historical data dominate the posterior distribution. Hence, the model can only be good if there is enough and appropriate historical data.

Concerning item b) of the first statement, a flat prior distribution for either the mean or the SD parameter in the new experiment results in a low power, and then the researcher will be unable to answer their biological question. Such a flat prior for the mean or SD may, for example, result from unaccounted heterogeneity in historical groups or experiments. Hence, it is important to reflect and best possible model any variables that may affect the experimental outcome. On the other hand, an over-optimistic assumption of the relatedness of the historical and the new data may lead to predictive distributions that are too informative and may lead to false positive controls.

The second statement asserts that, although the prediction for the SD does not influence the Bayes factor directly, it does still influence the posterior distribution for the new mean. Thus, also the model for the SD should be chosen carefully and not be neglected.

In summary, regarding the two statements above, they stress the importance of the model of the historical data and the model on the relatedness, that underlies the prediction. The availability and model of the historical data and the prediction are further discussed in Section 4.1.1.

To set up a decision rule that is based on the Bayes factor, a threshold has to be defined, that determines for which values of the Bayes factor the null hypothesis is rejected. In Table 9, a heuristic interpretation scheme for Bayes factors is presented [56, 119, 120]. In this thesis, the null hypothesis is rejected, if the Bayes factor BF_{10} is greater than three, indicating at least moderate evidence for the alternative according to Table 9.

Table 9: Heuristic categorization of evidence in a comparison of two hypothesis H_1 and H_0 using the Bayes factor BF_{10} based on Stefan et al. [56].

BF_{10}	More evidence for	Evidence category
$(100, \infty)$	H_1	Extreme
$(30, 100]$	H_1	Very strong
$(10, 30]$	H_1	Strong
$(3, 10]$	H_1	Moderate
$(1, 3]$	H_1	Anecdotal
$\{1\}$	None	No evidence
$[\frac{1}{3}, 1)$	H_0	Anecdotal
$[\frac{1}{10}, \frac{1}{3})$	H_0	Moderate
$[\frac{1}{30}, \frac{1}{10})$	H_0	Strong
$[\frac{1}{100}, \frac{1}{30})$	H_0	Very strong
$(-\infty, \frac{1}{100})$	H_0	Extreme

Using a threshold of moderate evidence from Table 9 corresponds to the decision rule

$$D_{\mathsf{BF}} = \mathbb{1}\{\mathsf{BF}_{10} > 3\}, \tag{24}$$

where $\mathbb{1}\{\mathsf{BF}_{10} > 3\}$ is the indicator function that takes the value one if the Bayes factor is greater than the threshold of three and the value zero if it is smaller or equal to three. As an alternative threshold the value ten, representing strong evidence according to Table 9, has been suggested in the context of sample size determination with Bayes analysis in psycho-linguistics [57]. In the application framework of translational animal experiments, hypothesis tests typically rather have an exploratory than a confirmatory character [7]. Hence, it may be considered more important to

not prevent potential treatments from being dismissed than in later stages of clinical trials. This can be realized by setting a lower threshold and already considering moderate evidence for a treatment candidate as being worth further investigation, although this choice may imply a higher probability of putting false hopes in treatment candidates that are in fact futile. The choice of the threshold and a more principle scheme for deriving decision rules based on Bayes factors is further discussed in Section 4.1.2.

Example 10: Bayes factor

In Example 9 the prior distribution for δ was determined as $\delta \sim \mathcal{N}(0, 2)$. The Bayes factor for testing the Bayesian hypothesis pair $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$ is calculated by the Savage-Dickey ratio (23) as ratio of the prior to the posterior distribution of δ under H_1 at $\delta_0 = 0$. Under H_1, δ may vary freely and is assigned the prior distribution $\delta \sim \mathcal{N}(0, 2)$. The height of the prior distribution at $\delta_0 = 0$ under H_1 is then $p_1(0|H_1) \approx 0.282$. The height of the posterior distribution under H_1 corresponds to the distribution $\delta|y \sim \mathcal{N}(0.727, 0.182)$ from Example 9 and is $p_1(0|y, H_1) \approx 0.219$. This gives $\text{BF}_{10} \approx 1.290$ (Figure 12). Since the Bayes factor is smaller than three (corresponding to moderate evidence for H_1 [56]), the null hypothesis cannot be rejected with the decision rule (24).

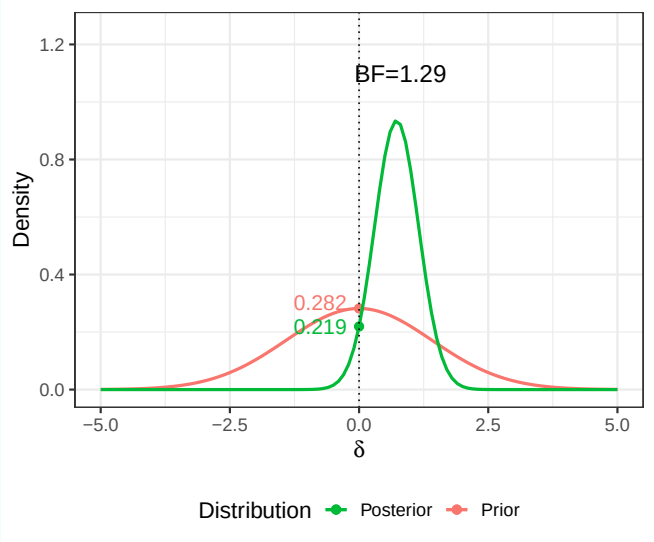


Figure 12: Probability density functions of the prior and posterior distribution of the mean difference δ and resulting Bayes factor BF_{10} (BF), as relative measure of evidence for the null hypothesis $H_0 : \delta = 0$ against the alternative hypothesis $H_1 : \delta \neq 0$.

Summary 2: Frequentist vs. Bayesian inference

	Frequentist analysis	**Bayesian analysis**
Informed by	Data alone	Data and prior distributions
Population/ "fixed" effects	Fixed values	Random variables
Estimation	Point estimates	Probability distributions
Uncertainty Intervals	Confidence interval	Credible interval
Interval boundaries	Random	Fixed
Interval interpretation	The probability that a confidence interval calculated in a random data set includes the true parameter value is $1-\alpha$.	Given the posterior credible interval of some data (and prior distribution), the probability that a random realization of the posterior distribution falls within the interval is $1-\alpha$.
Tests used here for testing $H_0 : \delta = 0$ (equal group means) vs. $H_1 : \delta \neq 0$	t-test, Welch test	Bayes factor indicating at least moderate evidence for H_1 [56], posterior credible interval for δ excluding 0
Small samples	Conclusions often rely on asymptotic arguments (large sample sizes needed)	Conclusions typically rely on exact distributions (or MCMC approximations), also appropriate for small samples
Computation	Usually less complex and less computer intensive	Usually more complex and computer intensive

2.3.2 Estimating power and determining sample size by simulation

The predicted SD from the historical data is used to calculate the necessary sample size with the frequentist t-test. The difference in the group means is defined as minimal (clinically or biologically) relevant mean difference $\delta^* > 0$. What is considered a minimal relevant mean difference depends on the experimental outcome and may be specified by the new experiment's researcher as an expert in the research field, as further discussed in Section 4.1.2. In the basic setting, equal group SDs are assumed in the new experiment (homoscedasticity). In this situation and since, in the frequentist approach, the historical information is only considered for the planning (as estimate for the SD in the new experiment) but not for the actual analysis of the new experiment, it is optimal to assign an equal proportion of the total sample size is allocated to the treatment groups of the new experiment. A total sample size N^* is considered large enough to detect true mean differences of $\delta(n^*)$ (or larger) if the t-test at level $\alpha = 5\%$ has a power of at least 80%. This gives group sample sizes of $n^* = N^*/2$ for the two treatment groups.

The frequentist sample size result is used as reference value for a possible sample size reduction with the Bayesian predictive approach. More specifically, in design simulations, it is investigated if, by using the predicted prior information in the Bayesian analysis of the new data, the sample size in the control group can be lowered without arriving at a power of below 80% or a type I error rate above 5% of the Bayesian test (24). The idea in the design simulation is to repeatedly sample animals from a design candidate (consisting of the parameters and the sample sizes candidates for the hypothetical new data and the prior distributions as in Section 2.2.3) and to evaluate the Bayesian test (24) for each simulated data set, as illustrated in Figure 13.

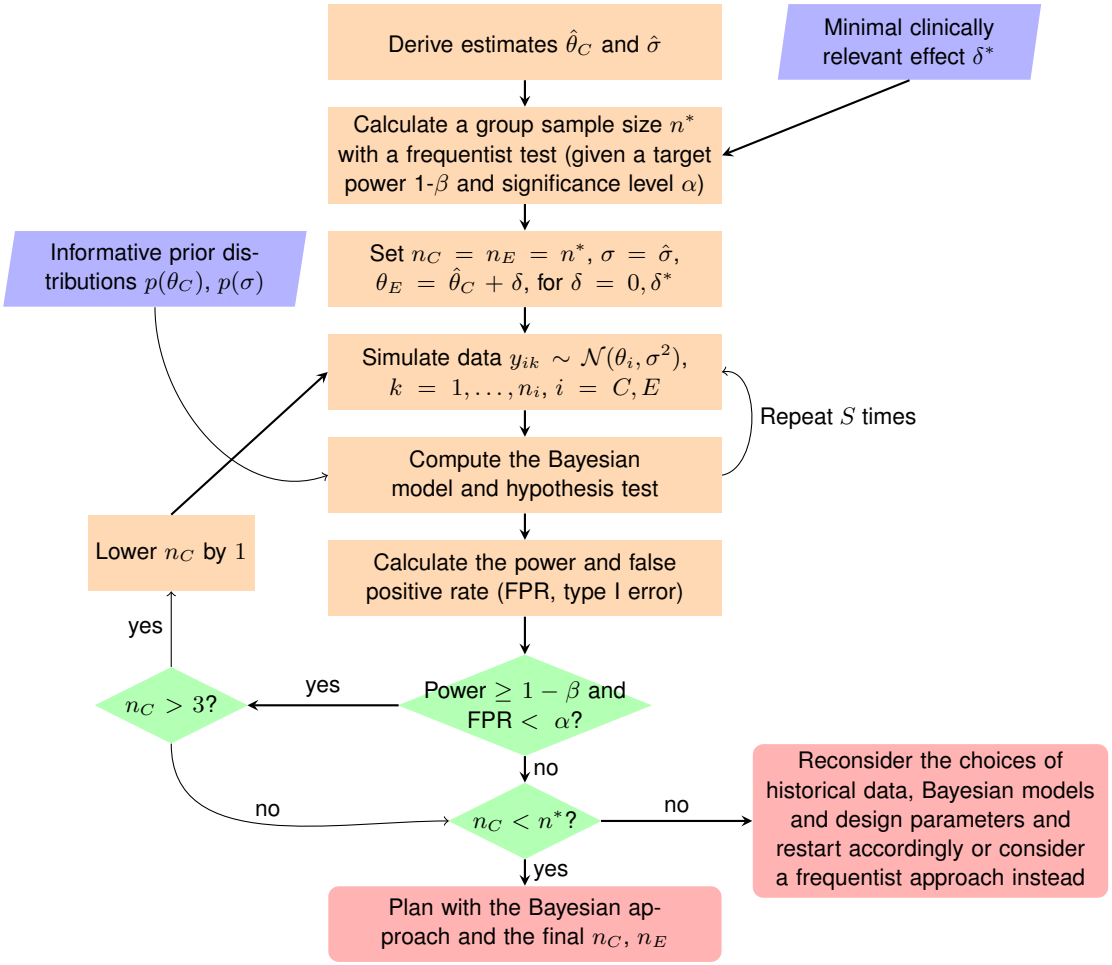


Figure 13: Design simulation part of the sample size planning procedure with Bayesian predictive approaches.

The power is calculated as percentage of data sets, where the null hypothesis of equal group means is rejected by the Bayesian test in all data sets that were simulated with a true effect of size $\delta^* \neq 0$. Accordingly, the type I error rate is calculated as percentage of data sets, where the null hypothesis of equal group means is rejected by the Bayesian test in all data sets that were simulated with a true effect of size 0. As minimal sample size in the new control group, three animals are considered, as recommended for exploratory animal experiments [6]. Alternatively, a minimal sample size of two could be chosen as the lowest number for which a SD can be computed [8]. If the Bayesian analysis results in a reduction of the sample size in the control group, while still maintaining a power of at least 80% and a type I error of at most 5%, the Bayesian test is preferred over the frequentist test and used for planning and analyzing the new experiment.

Example 11: Design simulation

The medians from the predicted prior distributions from the Bayesian hierarchical meta-analysis model of the historical data are used as basis for a frequentist sample size calculation with the two-sample t-test. The median of the predicted control mean is $\hat{\theta}_C \approx 0.13$ and the median of the residual SD $\hat{\sigma}_C \approx 0.98$. Under the assumption, that the SD in the experimental group is equal to the SD is in the control group, a common SD of $\hat{\sigma} = 0.98$ is used for the frequentist sample size calculation. Under the assumption that the animal researcher has specified a mean difference of $\delta^* = 1.5$ as minimal biologically or clinically relevant value, the mean for the experimental group is set to $\theta_E = 1.63$. With these choices, the resulting sample size per group that results from the frequentist two-sample t-test (calculated with the `power_t_test` function of the R `MESS` package [76] and a target power of 80% and a significance level of 5%) is

$$n^* = 8.$$

n^* is used as starting point to investigate how much the sample size can be reduced by including the predictive prior distributions in a Bayesian analysis of the new data. To this end, hypothetical data sets are simulated for the starting sample size $n^* = 8$ in both groups. The sampled data in the first simulation run are shown in Figure 14.

A Bayesian model is then fit to the simulated data using the predicted prior distributions from the historical data from Example 8 and MCMC computation. The prior distributions in the experimental group are chosen according to the considerations in Section 2.2.3: a normal UIP is chosen for θ_E with the same mean value as the prior for θ_C (as a conserva-

tive choice under the assumption, that no historical data is available for the experimental treatment in the new experiment's group). The SD parameter of the prior θ_E is set to the median of the predicted SD from the historical control data, which was 0.368. This leads to a wider and thus less informative prior distribution for θ_E than for θ_C.

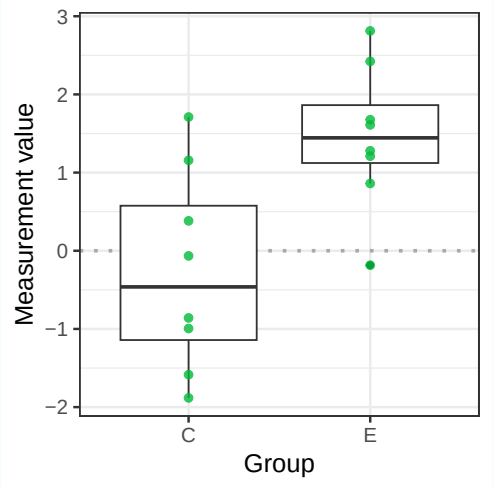


Figure 14: Sampled data in the first run of the design simulation, representing hypothetical measurements in the control (C) and experimental (E) group in the new experiment.

The Bayesian model estimates posterior distributions for the group means and the common residual SD, which are shown together with their prior distributions in Figure 15. The posterior distributions of the group means are used to calculate the mean difference δ, which is then used for hypothesis testing with the Bayes factor and the hypothesis from Example 10. In Figure 16 the Bayes factor (as ratio of the prior to the posterior density of δ at the value zero) is shown. The value of the Bayes factor is greater than the value three (which indicates moderate evidence for $H_1 : \delta \neq 0$ vs. $H_0 : \delta = 0$ [56]), so H_0 is rejected with this data. The steps of data simulation, Bayesian model fitting, and hypothesis testing are repeated for further S times (e.g. $S = 10000$). Then, the power and false positive rate (FPR, type I error) can be calculated as proportion of rejections of the null hypothesis in the S simulations. If the power is large enough ($\geq 80\%$) and the FPR is small enough ($< 5\%$), the sample size in the control group is lowered by one, leading to a new sample size of n_C in the next set of simulations. The procedure stops if the power is below 80% or the FPR is $\geq 5\%$ and then additional design quality criteria are considered, as introduced in the next section. A final choice on a design is made, with results shown in Section 3.4.

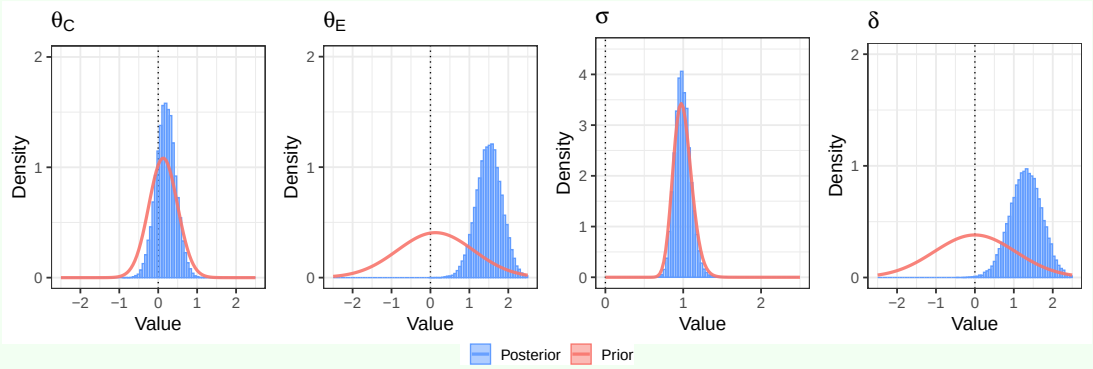


Figure 15: Probability density functions (red lines) of the prior distributions for the group means θ_C and θ_E, the common residual SD σ and the mean difference δ and histograms (blue) representing the relative frequencies of the corresponding posterior MCMC draws.

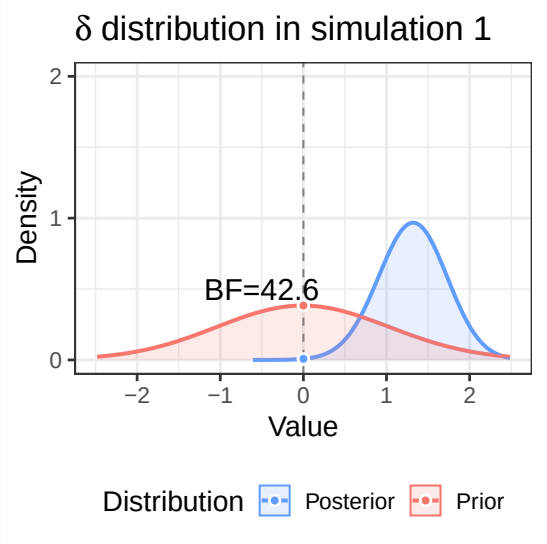


Figure 16: Bayes factor as ratio of the prior to the posterior density of the mean difference δ in the first run of the design simulation.

2.3.3 Additional design quality criteria

To get a more comprehensive understanding of the impact of a design candidate on the biological conclusion concerning the treatment effect and to supplement the goal of statistical power by further aspects, additional model characteristics are examined [1, 121]. These characteristics refer, e.g., to long-term properties for estimates $\hat{\delta}$ of the posterior treatment difference compared to a fixed true effect δ^*. In the Bayesian setting, where the mean difference is modeled as a random variable δ, a point estimator can be derived as the mean or median of the posterior distribution $p(\delta|y)$. One quality aspect of a point estimator $\hat{\delta}$ of the true difference δ^* is addressed by its *bias*:

$$\text{Bias}(\hat{\delta}) = \text{E}(\hat{\delta}) - \delta^*. \tag{25}$$

The bias refers to the expected amount by which the point estimator deviates from its estimand (the true value δ^*). An unbiased estimator has a bias of zero. In the frequentist case and with normally distributed data, unbiased estimates of the treatment group means can be derived as the arithmetic sample means (4).

Gelman and Tuerlinckx [122] argue to consider the probabilities conditional on significance claims or "claims with confidence" because these claims are most critical from a practical perspective. In translational animal experiments, significance claims correspond to the experiments where treatment candidates are claimed to actually have an effect on the experimental outcome (that represents the pathological phenotype) and may thus be worth further investigation. As suggested by Gelman and Carlin [12, 122], the conditional *type S (sign) error rate* and the conditional *exaggeration factor* are examined as additional design characteristics. The conditional type S error rate is the probability that the estimate has the wrong sign, given that is statistically significant, and the conditional exaggeration factor is calculated from the type M (magnitude) errors as the ratio by how much the absolute value of the effect estimate is greater than the absolute value of the true effect, given that it is statistically significant:

$$s(\hat{\delta}) = \text{P}\left(\text{sign}(\hat{\delta}) \neq \text{sign}(\delta^*)|D(y) = 1\right)$$
$$m(\hat{\delta}) = \text{E}\left(\frac{|\hat{\delta}|}{|\delta^*|}|D(y) = 1\right)$$

$D(y)$ denotes the decision function which is used for hypothesis testing. The decision functions are here the two-sample t-test (3) or Welch test (6) for the frequentist designs and the Bayes factor test and credible interval based test from equations (24) and (21) for the Bayesian case. For a collection of estimators $\hat{\delta}_s$, that represents the calculated values of an estimator $\hat{\delta}$ in S random data sets $y_s = \{y_{sC1}, \ldots, y_{sCn_C}, y_{sE1}, \ldots, y_{sEn_E}\}$, $s = 1, \ldots, S$, (representing the control and experimental animals that are simulated as hypothetical data for the new experiment with true mean difference of $\delta_s^* \neq 0$ $s = 1, \ldots, S$), the sample type S error rate and exaggeration factor of $\hat{\delta}$ are calculated as averages in those data sets $\tilde{s} = 1, \ldots, \tilde{S}$, where the corresponding hypothesis test $D(y_{\tilde{s}})$ was evaluated to one (rejection of H_0):

$$\hat{s}(\hat{\delta}) = \frac{1}{\tilde{S}} \sum_{\tilde{s}=1}^{S} \mathbb{1}\{\text{sign}(\hat{\delta}_{\tilde{s}}) \neq \text{sign}(\delta_{\tilde{s}}^*)\} \tag{26}$$

$$\hat{m}(\hat{\delta}) = \frac{1}{\tilde{S}} \sum_{\tilde{s}=1}^{S} \frac{|\hat{\delta}_{\tilde{s}}|}{|\delta_{\tilde{s}}^*|} \tag{27}$$

$\mathbb{1}\{\text{sign}(\hat{\delta}_{\tilde{s}}) \neq \text{sign}(\delta_{\tilde{s}}^*)\}$ denotes the indicator function, that takes the value one if the sign of the estimated mean difference in a simulated data set is equal to the sign of the true mean difference and zero if the signs are different.

Lu et al. [123] derive closed-form expressions for the type S error rate and the exaggeration factor in a frequentist setting and consider the two quantities as functions of the significance level and standardized effect size (as treatment difference divided by the standard deviation of the data). The closed-form expressions show that if the standardized effect size increases, both the type S error and the exaggeration factor monotonically decrease. In the limiting case of an infinite large effect size, the type S error rate approaches zero (since the estimates are all well above or below zero) and the exaggeration factor approaches one (because the estimate can no longer overestimate the effect size). On the other hand, when the standardized effect size approaches zero, the type S error rate approaches 50%, since the sign of the estimate then corresponds to the outcome of a fair coin flip. The exaggeration factor approaches infinity in the situation of the standardized effect size approaching zero since the estimator will always overestimate the infinitely small effect size [123]. This implies that it is for large treatment effects likely, that when the hypothesis test rejects the null hypothesis of no effect, the treatment will truly have a non-zero effect and the estimate of the effect will probably not be over-optimistic of the effect. In contrast, problems occur with small effect sizes, since those will often lead to over-optimistic estimates and further investigation of in fact futile treatment candidates. Small effect sizes may occur either due

to small differences in the treatment means or due to large standard deviations of the data. In the Bayesian setting, the values of the type S error rate, the type M error rate and the exaggeration factor depend on the choice of prior distribution, but in general, the values can be high in noisy data, that is characterized by a large standard error [124, Section 4.4]. In practice, type M error rates are of special importance, in particular with regard to the "statistical significance filter": in small sample experience, the standard errors of the estimates tend to be large, and with larger standard errors, also the effect estimates have to be large to achieve statistical significance. This implies type M errors and high exaggeration factors in the published literature that prioritizes experiments with significance claims [125], [124, Section 4.4]. When designing a new animal experiment, it is thus important to get a feeling of how reliable the claims will be. In addition, the investigation of type S error rates and exaggeration factors can also help in a retrospective design analysis to judge the reliability of the results [12], as further discussed in Section 4.1.2.

As further characteristic, the root mean square error (RMSE) as square root of the mean square error (MSE) is examined:

$$\text{MSE}(\hat{\delta}) = E\left((\hat{\delta} - \delta^*)^2\right)$$

The MSE can be written as the sum of the bias of the estimator (as defined in (25)) plus the variance of the estimator. For an unbiased (with Bias equal to zero), the sum reduces to the variance of the estimator. The RMSE has the interpretational advantage of having the same unit as the estimand. For a collection of estimators $\hat{\delta}_s$, that is calculated in the above defined simulated data y_s, $s = 1, \ldots, S$, for the new animal experiment, the sample RMSE can be calculated as

$$\widehat{\text{RMSE}}(\hat{\delta}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\delta}_s - \delta_s^*)^2}. \tag{28}$$

Both MSE and RMSE are always non-negative and generally lower values are preferable, but there is no universal interpretation scheme since the measure depends on the scale of the data and the application context. For example, if the experimental outcome can take values in the range from zero to 100, then a RMSE 0.5 seems rather small in most practical applications. However, if the experimental outcome can take values in the range from zero to two, then a RMSE 0.5 seems rather large. The MSE and RMSE can be calculated from simulated datasets (representing the new experiments) as the arithmetic mean of the squared distances of the true values of δ_s^*, that were used for the data simulation, and their estimates $\hat{\delta}_s$ in the simulated data sets, $s = 1, \ldots, S$. By definition, each error goes into the sum by its square, so the quantities

are sensitive to extreme estimates. A measure that is less sensitive to extreme estimates is the mean absolute error of the estimator, which is defined as the average of the absolute differences $\hat{\delta}_s - \delta_s^*,\ s = 1, \ldots, S$. While in the MSE and RMSE the differences are squares, each error goes in the mean absolute error in proportion to its absolute value. In this thesis, the RMSE is used for design evaluation instead of the mean absolute error, since large errors are regarded as particularly undesirable.

Additionally, the average widths of the frequentist confidence intervals and the Bayesian credible intervals $\text{CI}_s = (\text{CI}_{s,\text{upper}}, \text{CI}_{s,\text{lower}})$ from (16) and (17) are calculated for the S simulated data sets $s = 1, \ldots, S$ as $\frac{1}{S}\sum_{s=1}^{S}(\text{CI}_{s,\text{upper}} - \text{CI}_{s,\text{lower}})$. Narrower confidence intervals imply more reliable estimates and generally make it easier to derive precise statements on the effect of interest. By definition, frequentist confidence intervals are set up to include the true estimate with a probability that is defined by the anticipated coverage rate (typically 95%). However, when the data in the new experiment deviate from the probability model, that is assumed for constructing the interval, then the interval may be too narrow (or too wide) and the actual coverage probability may be below (or above) the nominal level of 95%. The interval coverage rate is defined as the probability of the confidence or credible interval including the true effect δ^* and is calculated in the simulated data sets as binomial estimator $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\{\delta_s^* \in \text{CI}_s\}$.

The overall sample size planning procedure with Bayesian predictive approaches, consisting of both the first step of fitting a Bayesian model in the historical data and making predictions for the new experiment and the second step of using these predictions in the design simulation and hypothesis testing for planning the new experiment, is graphically summarized in Figure 17. The procedure will be applied to a real-world and a fictive animal experiment example that are introduced in Section 2.4.

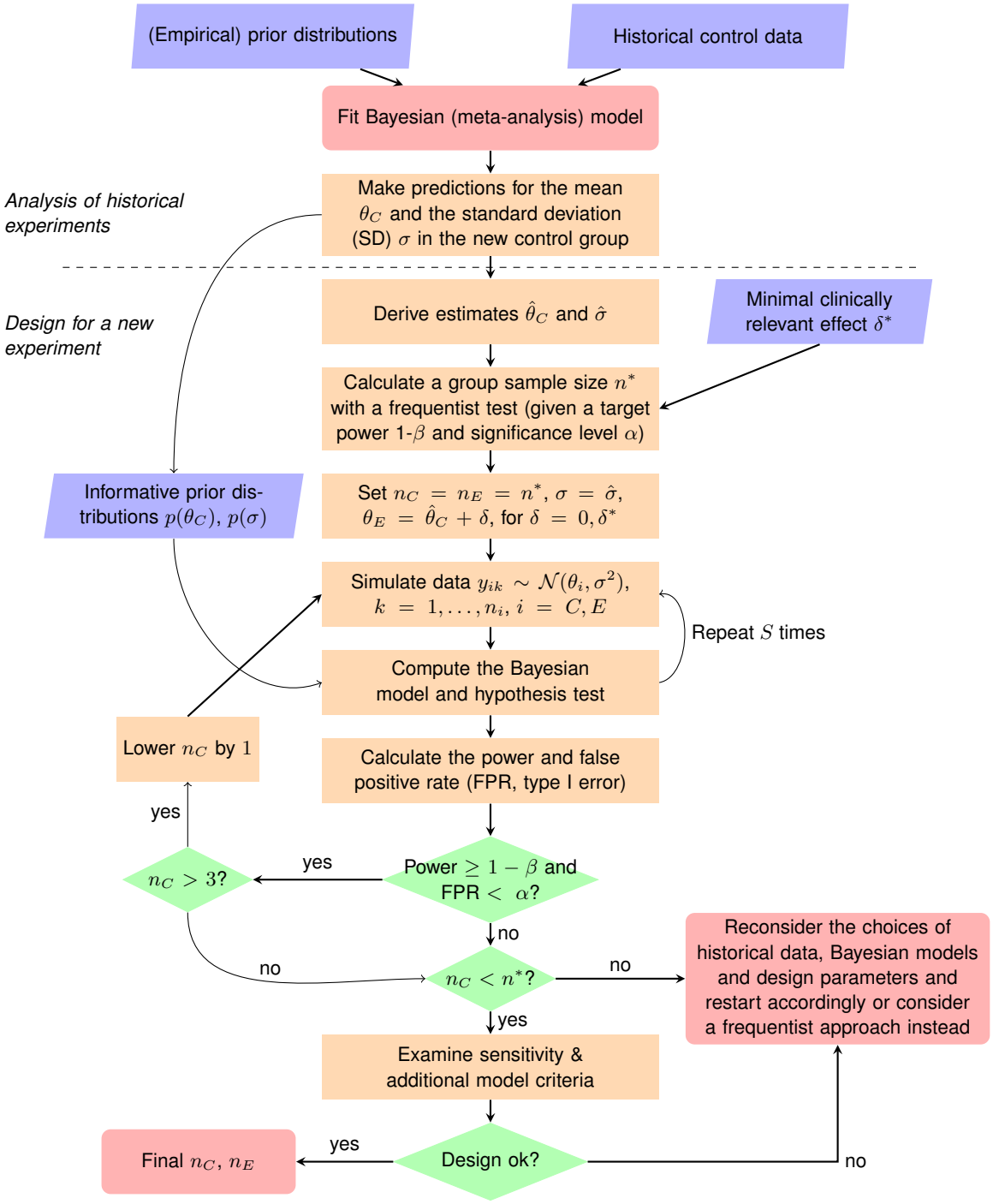


Figure 17: Sample size planning procedure with Bayesian predictive approaches.

Summary 3: Sample size determination by simulation based design analysis

- Classically, sample sizes are set up to obtain a target prospective power $1-\beta = 80\%$ for detecting effects δ greater than a biologically or clinically relevant mean difference δ^*, while limiting the probability of type I errors at level $\alpha = 5\%$. In standard applications, then, the sample size is calculated by solving power (in)equalities or iteratively calculating the associated power for a sequence of increasing sample sizes, until the target power is reached.
- Setting an optimal sample size in simulation based design analysis represents a more general approach to sample size calculation and can be applied with arbitrary (Bayesian) models.
- In simulation based design analysis, hypothetical data, consisting of the control and experimental animals in the new experiment, are repeatedly simulated under various sample sizes and assumptions on the groups' mean parameters and on their variances.
- In the simulated data, error rates of predefined hypothesis tests can be calculated. Analogue to frequentist error probabilities, the power can be calculated as probability of correctly rejecting the null hypothesis H_0, when the data were simulated from the model under the alternative hypothesis H_1, and the type I error rate as probability of falsely rejecting H_0, when the data was actually simulated from the model under H_0.
- An optimal sample size can then be chosen as the minimal sample size, that meets a predefined statistical goal (e.g., a power of at least 80% of a hypothesis test).
- Bayesian models are typically more complex and there is no common census what should be the statistical goal of the experimental planning (as an equivalent to frequentist power). Analyzing designs solely based on power and type I error rates bears the risk to miss other important design properties, so the analysis of additional aspects (precision, bias, exaggeration ratios and the robustness of the conclusions to violations of model assumptions) is recommended.

2.4 Real-world example "bone quality under postmenopausal osteoporosis"

The sample size determination procedure, which was introduced in the previous section, was applied to the planning of a real-world animal experiment from translational research, as introduced in my previous contribution [1]. The planned animal experiment aims to examine the role of the complement component 5a receptor 1 (C5aR1) on bone quality under the pathological condition of postmenopausal osteoporosis. The pathological condition of postmenopausal osteoporosis shall be examined in mice having undergone ovariectomy. Ovariectomy is associated with a deficiency in the ovarian hormone estrogen and bone loss [126], whose acute effects can optimally be investigated in ovariectomized mice [127]. Estrogen regulates the bone remodeling process and an estrogen deficiency increases the osteoclast activity and bone resorption of osteoclasts [128]. Osteoblasts are cells that are responsible for the formation of bone tissue, while osteoclasts, as their counterparts, cause the resorption of bone tissue. Bone remodeling requires a proper balance between osteoclasts and osteoblasts to maintain bone mass [129]. The peptide C5a is a product of the complement factor C5 [130] which affects the bone remodeling process [131–133]. In particular, reduced numbers of osteoclasts and increased bone mass were observed in a knockout mouse model with global C5aR1 knockout [132].

The aim of the planned experiment is to investigate how and if a specific knockout of the C5aR1 in osteoblasts and osteoclasts affects the bone quality in ovariectomized mice. The inhibition in osteoblasts is realized by a Cre-recombinase-specific knockout of the transcription factor runt-related transcription factor 2, which controls osteoblast differentiation, while the osteoclast inhibition is realized by a knockout of the lysozyme M domain, with the background that lysozyme M-positive macrophages are precursor cells of osteoclasts [134]. The bone quality shall be measured by two main experimental outcomes, which are the bending stiffness in a biomechanical test and the relative bone volume (bone volume to tissue volume (BV/TV)) in a micro-computed tomography (micro-CT) scan. This experiment will be referred to as *C5aR1 experiment* in the remainder of this thesis. The planned experiment shall consist of female mice with a C57BL/6J background (as a common inbred strain in laboratory mice [135]), which are of age twenty weeks at the time when the outcomes are measured. If the experimental treatment (C5aR1 knockout) has an effect on the outcomes in the planned experiment, and if such an effect can be confirmed in experiments with additional rodent strains and clinical trials in humans [127], then the knockout may serve as a basis for the development of treatments for the pathological phenotype of postmenopausal osteoporosis in humans.

2.4.1 Historical data

In the following section, the available historical data, that can be used for planning the C5aR1 experiment is introduced. In my previous contribution [1] internal historical data from the experimental researcher, who was planning the new experiment, had been introduced for the outcome BV/TV. These internal data had been supplemented with external phenotyping data from the Mouse Phenome Database (MPD) of the Jackson Laboratory [58, 136]. The MPD is an open-access biological database of phenotyping mouse data on an individual-animal level that come in a tidy standardized format. Besides the data, the MPD provides researchers' with experiment protocols describing the experiments' procedures, and statistical tools for data analysis. Data in the MPD is endowed with the Mammalian Phenotype Ontology [137] as common phenotype terms, definitions, and synonyms that can be used to describe mammalian phenotypes. Additional biological phenotypic databases and publications on `pubmed` (https://www.ncbi.nlm.nih.gov/) were considered, but no additional individual-level data were found for the experimental outcomes of interest. The benefits and challenges in using data from the MPD are further discussed in my previous contribution [1] and in Section 4.1.1.

As stated in the introduction, my previous contribution [1] introduces aspects of fitting Bayesian hierarchical or meta-analysis models to historical data in translational animal experiments and using them as predictive distribution for the investigation of different design properties in a simulation approach for the new experiment. In this thesis, these ideas are generalized to alternative prediction models and are summarized to a workflow for sample size planning, as introduced in Section 2.2.3. The exemplary design planning and analysis for the outcome BV/TV from my previous contribution [1] is complemented with an application of the sample size planning procedure for the outcome bending stiffness in the C5aR1 experiment and with fictive historical data, that illustrate problems in classical sample size calculation based on a single small historical experiment.

External data for the outcome bending stiffness

The search term "biomechanical test" led to wild type data from four animal experiments, as presented in Table 11. The animals from the experiments of Donahue et al. [138] and Beamer [139] are mice of the same age as the youngest mice in the internal historical data. The mice in the experiments of Wergedal et al. [140] are slightly younger than the internal historical mice, and the mice in the experiment of Burke et al. [141] are notably older than the internal mice. The mice of Donahue et al. consist of ten different inbred strains. The mice of Beamer [139], Wergedal et al. [140], and Burke et al. [141] are second filial (F2) mice as descendants of hybrid first filial (F1)

mice. A hybrid cross in turn results from a cross of two different inbred strains [142]. With this wide range of strains, it is expected that the data gives a representative picture of the distribution of the outcomes in mice in general. The relevant measurements for the biomechanical tests are presented in Table 12. They are measurements from a 3-point bending test as described by Goodyear and Aspden [143]. As part of the MPD, they are endowed with a unique vocabulary or ontology.

Table 11: Meta data (publication year, sex (m: male, f: female), animal age (in weeks)), animal strains and average number of animals per subgroup ($\overline{n}_g$, $g = 1, \ldots, 4$) (grouped by sex and strain) of the historical experiments in mice from the Mouse Phenome Database (MPD) resulting for the search term "biomechanical test".

Experiment	Year	Sex	Age	Animal strains	$\overline{n}_g$
Donahue [138]	2004	m, w	16	129S1/SvImJ, BALB/cJ, C3H/HeJ, C57BL/6J, C58/J, FVB/NJ, NZB/BINJ, PWK/PhJ, SJL/J, SWR/J	10
Beamer [139]	2001	w	16	(B6xC3H) F1 x (B6xC3H) F1	999
Wergedal [140]	2006	w	10	(NZB x RF) F1 x (NZB x RF) F1	661
Burke [141]	2012	w	80	(BALBxB6) F1 x (C3HxDBA/2) F1	505

Table 12: Variables with their units from four data sets of the MPD that refer to the biomechanical test. N: Newton. mm: millimeter.

Experiment	Variable	Description	Unit
Donahue	femur_peak_load	maximum load on load-displacement curve	N
Donahue	femur_break_load	load at specimen failure (break point)	N
Donahue	femur_peak_extension	displacement at peak load	mm
Donahue	ultimate_strength	estimated ultimate strength	N
Donahue	femur_stiffness	slope of force-displacement curve in linear region	N/mm
Donahue	elastic_modulus	estimated elastic modulus	N/mm^2
Beamer	femur_break_load	load at specimen failure (break point)	N
Beamer	femur_stiffness	stiffness	$N \cdot mm$
Beamer	femur_tough	elastic modulus (toughness)	N/mm^2
Beamer	femur_stress	ultimate stress	N
Beamer	femur_energy	energy to failure	$N \cdot mm$
Beamer	femur_inertia	cross-sectional moment of inertia	mm^4
Wergedal	P	maximum (ultimate) load	N
Burke	Vert_Ult_Load	vertebra mechanical load at breaking	N

To understand if and how differently defined measurements are related or may be transformed to be comparable, an understanding of biomechanical tests is necessary. Furthermore, an understanding of the relationship between measurements from different experiments is important to set up a correct statistical model of historical bending stiffness data and for the prediction model of the bending stiffness in the new experiment. What complicates the situation is that researchers may refer to different definitions of bending stiffness: extrinsic and intrinsic stiffness

[143]. In a bending test, a bone is attached with both ends to a tray and loaded with weight until it breaks. The bending, as displacement, is recorded as a function of the applied load in a load-displacement curve. The load-displacement curve is used to represent extrinsic properties of the bone, while intrinsic properties are represented in the stress-strain curve, as shown in Figure 18. More specifically, the intrinsic stiffness is the elasticity modulus $E = \frac{\sigma}{\epsilon}$, defined as the slope of the linear region of the stress-strain curve. In the internal data, "stiffness" refers to extrinsic stiffness, which is the slope of the linear area in the stress-strain curve or calculated from the intrinsic stiffness as product of the intrinsic stiffness and the moment of inertia, I, of the bone cross-section:

$$\text{(Extrinsic bending) stiffness EI} = E \cdot I \tag{29}$$

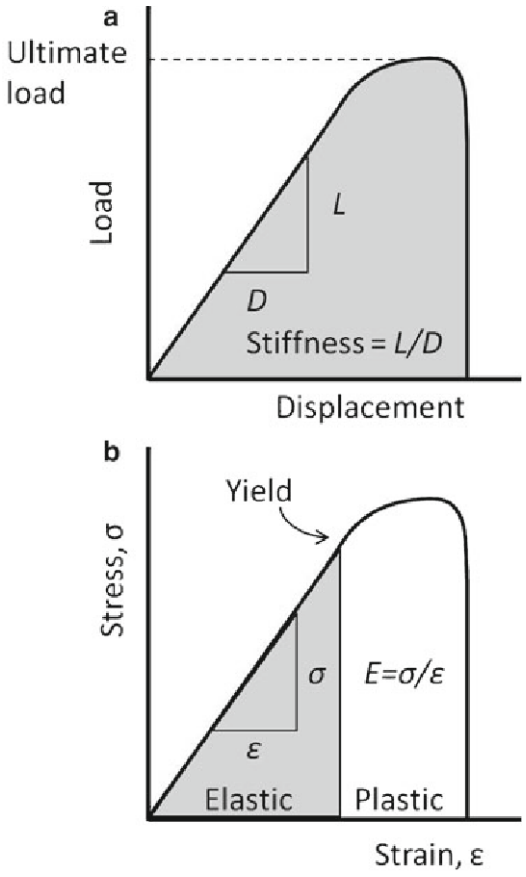


Figure 18: a) Load-displacement curve for representation of extrinsic properties like stiffness (slope) and ultimate load. b) Stress-strain curve for representation of intrinsic properties like the elasticity modulus and yield. Figure adopted from [143] with permission of SPRINGER PROTOCOLS.

However, since the bone is not always in the middle of the two support points of the testing machine, a modified formula was used for calculating the extrinsic stiffness in the internal data:

$$\widetilde{\mathsf{EI}} = \frac{\mathsf{EI}}{3l} l_1^2 l_2^2 \tag{30}$$

Here, l_1 and l_2 are the distances between the loading vector and the proximal and distal attachment points and $l = l_1 + l_2$ is the distance between the support points.

The product EI in (29) can be calculated from the MPD data by multiplication of the variables femur_tough and femur_inertia. However, the cross-sectional moment of inertia (I) is not given for the Donahue experiment and the necessary information for transforming the product EI to the scale of the extrinsic stiffness variable in the internal data is missing. As an alternative outcome, stiffness could be considered in terms of the slope of the load-deformation curve (variable "femur_stiffness" in Table 12 of the MPD variables), instead of the extrinsic stiffness definition in (30). However, "femur stiffness" is measured in N/mm (Newton/millimeter) in Donahue's experiment while it is measured in $N \cdot mm$ in Beamer's experiment, also the measurement descriptions and identifiers differ. If the comparison of measurements from different experiments raises doubts about the comparability of the measurements, then the research protocols and potential reasons for the deviations should be further investigated. This may lead to the decision to exclude some historical data from the analysis.

External data for the outcome relative bone volume

The search term "BV/TV" on the MPD web page led to additional data from 31 strains of Collaborative Cross (CC) wild type mice from an experiment of Levy et al. [144]. CC mice are bred from genetically divergent strains and are thus represent a high degree of the allelic diversity in the mouse genome [145]. They are therefore also assumed to give a broad picture of plausible values for relative bone volumes.

2.4.2 Fitted Bayesian models

In the planned experiment, the mice shall undergo ovariectomy and the comparisons of interest are those between the ovariectomized animals with and without a knockout in the C5aR1. Since ovariectomy is expected to have an impact on the outcomes bending stiffness and relative bone volume, the MPD mice were not used for the prediction of the control mean and SD. Alternative uses of such historical control data, that do not meet the definition of the control group in the

new experiment, but still provide relevant information on the experimental outcome of interest, are discussed in my previous contribution [1] and in Section 4.1.1. In my previous contribution [1], the results from fitting a Bayesian hierarchical (meta-analysis) model to both internal data and external data from the MPD were presented. Due to the differences in the control groups from the MPD and the new experiment, the mean of the SD of the new data were predicted from the internal data only. The model could be used to define a range of plausible effect sizes for the design of the C5aR1 experiment. In the first part of this section, an alternative historical data model is presented, that uses only the internal historical data. In addition, a model for the historical bending stiffness data is presented. In the second part of the section, the models for simulating the hypothetical new data in the planned experiment and the Bayesian prediction models for the model parameters are presented. The choice of prior distributions was based on the considerations from Section 2.2.3 and the considerations from my previous contribution [1].

Bayesian models for the historical data

Bending stiffness

For the outcome bending stiffness, the internal data consisted of three age groups (16, 20, and 52 weeks). Since, by definition, bending stiffness is restricted to the positive numerical range, the values are transformed to a square root scale to better resemble the samples of a normal distribution. As an alternative transformation, a logarithmic transformation could have been used after manipulating the data by adding a small positive constant to those values where a bending stiffness of zero was observed. However, this did not lead to a better resemblance to the normal distribution, and hence the square root transformation was preferred to avoid additional manipulation of the original values. Since the number of distinct observed ages was only three and since the data suggested a non-linear effect of age, age was modeled as a categorical variable, which allows also non-linear effects, and the model was fit to the internal data of the new experiment's researcher was

$$y_k = \mu + \beta_{\mathsf{Age}_{16}} x_{\mathsf{Age}_{16;k}} + \beta_{\mathsf{Age}_{52}} x_{\mathsf{Age}_{52;k}} + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2), \tag{31}$$

for animal $k = 1, \ldots, N$, where $N = 21$ is the total sample size of the internal, historical, ovariectomized animals. The intercept μ represents the reference level of the variable age of 20 weeks, which is also the age of data in the new experiment. The coefficients x_{Age_l} for week $l = 16, 52$ represent the expected change in the measurements of the bending stiffness in those animals

that are either 16 or 52 weeks old over those animals, that are 20 weeks old. The corresponding dummy-coded age variables are

$$x_{\text{Age}_{l;k}} = \begin{cases} 1, & \text{if animal } k \text{ is } l \text{ weeks old,} \quad l = 16, 52,\ k = 1, \ldots, 21, \\ 0, & \text{else.} \end{cases}$$

In a Bayesian version of the data model (31) prior distributions have to be set for all mode parameters (μ, $\beta_{\text{Age}_{16}}$, $\beta_{\text{Age}_{52}}$, σ). As illustrated in Section 2.2.3, the prior distributions could be set in an empirical Bayes approach based on properties of the modeled data themselves. , as suggested in Section 2.2.3. However, to avoid using the same data twice in the Bayesian model (once for the definition of the prior distributions and once as actual observations), additional internal historical data of wild type mice is used for setting up the prior distributions. Using independent data for setting up the prior and building the model reflects the Bayesian thinking of the prior as a priori knowledge better [146]. On the one hand, including a larger amount and variety of information may increase the representativeness of the model [147] and lessen the impact of single extreme observations. On the other hand, setting up prior distributions by observed bending stiffness measurements from wild type animals, which may be quite different from the measurements in the modeled ovariectomized animals, may also result in prior-data conflicts and unintentional sensitivity of the posterior distributions and the medical conclusion on the choice of prior distribution [38]. For these reasons, it is important to be aware of the influence of the prior distribution on the posterior conclusions. The chosen prior distributions and the topic of prior-data conflicts are further discussed in Section 4.1.1. The here chosen prior distributions for the bending stiffness model (31) are presented in Table 13.

Table 13: Prior distributions for the Bayesian regression model of the historical data of the square root bending stiffness in ovariectomized mice. Mean and SD: expected value and standard deviation of the distributions. Median and 95% Interval: 0.5, 0.025 and 0.975 quantiles of the distributions.

Parameter	Distribution	Family	Mean	SD	Median	95% Interval
μ	$\mathcal{N}(46, 4.9^2)$	Normal	46.0	4.90	46.0	(36.4,55.6)
β_{Age_l}	$\mathcal{U}(-\infty, \infty)$	Uniform				
σ	$\mathcal{HN}(4.9^2)$	Half-Normal	3.9	2.95	3.3	(0.2,11.0)

According to the considerations from Section 2.2.3, the intercept μ is modeled by a normal prior distribution and the SD σ by a half-normal distribution (that is suited as weakly informative prior distribution and can also be used in the situation of few historical observations). The parameters

of the normal distribution for μ are set according to the estimated mean and SD of the internal wild type mice. The scale parameter of the half-normal distribution for the SD is also set to the estimated wild type SD. This choice leads to a prior distribution, that has about 80% of its probability mass in the interval $(0.6, 7.6)$ and 95% in the interval $(0.2, 11)$, which seem like plausible ranges in light of the data of the mice without ovariectomy (which are presented in Chapter 3).

Relative bone volume

For the outcome relative bone volume, the internal data consisted of female ovariectomized mice of only one age (23 weeks) [1]. Like the bending stiffness measurements, the BV/TV measurements can, by definition, only take positive values. In the historical data, most of the measurements were close to zero (but none were actually zero, unlike the observations for the bending stiffness), while some of the measurements were very large. As a transformation of strictly positive values to the real numbers, a logarithmic transformation of the BV/TV measurements was considered. Then, an intercept-only model was fit to the internal historical data as

$$y_k = \mu + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2) \tag{32}$$

for animal $k = 1, \ldots, N$, where $N = 12$ was the total sample size of the internal ovariectomized mice. Since for this outcome, the external MPD data seemed to be on a similar scale as the internal data, the MPD data of female wild type mice was used to specify prior distributions for the model of the internal data. For the intercept parameter μ a normal distribution was used as prior distribution with mean and standard deviation parameters that equal approximately the arithmetic mean and empirical standard deviation of the female MPD animals. The residual standard deviation was given a half-normal prior with scale parameter according to the empirical standard deviation of the female MPD animals. The chosen prior distributions are summarized in Table 14.

Table 14: Summary of the prior distributions for the Bayesian regression model of the historical data of the logarithmic relative bone volume in ovariectomized mice. Mean and SD: expected value and standard deviation of the distributions. Median and 95$\%$ Interval: 0.5, 0.025 and 0.975 quantiles of the distributions.

Parameter	Distribution	Family	Mean	SD	Median	95$\%$ Interval
μ	$\mathcal{N}(2, 0.7^2)$	Normal	2.0	0.70	2.0	(0.63,3.37)
σ	$\mathcal{HN}(0.7^2)$	Half-Normal	0.6	0.42	0.5	(0.02,1.57)

Bayesian models and sample size calculation for the new experiment

Simulation-based design analysis was performed as described in Section 2.3 and Figure 17: first, the sample size n^* was calculated, that would be necessary to detect effects of δ^* with a power of at least 80% in a frequentist two-sample t-test at a two-sided significance level of $\alpha = 5\%$. The value δ^* represents a context-specific minimal (biologically or clinically) relevant effect as mean difference of the treatment groups. In practice, δ^* is specified by the experimental researcher based on available knowledge about the experimental outcome and on expectations on the experimental treatment. Then, it was checked if and how much the frequentist sample size n^* could be reduced in the control group by using the predictions from the historical data as prior distributions in a Bayesian analysis. For this purpose, hypothetical control and experimental animals for the new experiment were repeatedly simulated as described in Section 2.3.

Under the assumption of homoscedasticity (equal variances) in the two treatment groups, the data were simulated as

$$y^*_{ik,s} = \theta^*_i + \epsilon^*_{k,s}, \quad \epsilon^*_{k,s} \sim \mathcal{N}(0, (\sigma^*)^2) \tag{33}$$

for animal $k = 1, \ldots, n_i$ in group $i =$ C,E of the simulated data set $s = 1, \ldots, 10000$. The sample size n_E in the experimental group was kept at the value n^* from the frequentist calculation. This is because in the experimental group no historical information can substitute the corresponding information in the new experiment, that is required for answering the research question of whether a knockout in the C5aR1 changes the average bending stiffness. In contrast, in the control group, the frequentist sample size of $n_C = n^*$ is iteratively lowered to investigate if including the predictive information from the historical data can lead to a lower sample size in the Bayesian analysis.

As a further sensitivity analysis, the power and type I error rate, which were obtained from a Bayesian analysis with informative prior distributions for the mean and standard deviation parameter (obtained from predictions from the historical data model) were compared to the power and type I error rate under different options for the prior distributions of those parameters. Firstly, the power with the informative predictive prior distributions for the mean and variance parameter in the control group was compared to the power with a less informative unit information prior (UIP) for the control group mean (as defined in Section 2.2.3). For the mean in the experimental group, a UIP was used in each set-up, since this corresponds to the initial choice of the prior, without additional historical information. Moreover, designs were analyzed where the residual SD was modeled as a fixed value instead of assigning it a prior distribution and modeling it as a random variable. The intention behind this design is to investigate if there is a benefit in modeling the uncertainty in the

SD in a completely Bayesian design instead of a partly Bayesian design, as described in Section 2.2.3 and Example 4. The fixed value in the partly Bayesian design was set equal to the median of the predicted residual standard deviation from the historical data model. For investigation of the heteroscedastic model (34), two independent prior distributions were defined for the control and the experimental group. The heteroscedastic treatment groups were simulated as

$$\begin{aligned} &y^*_{ik,s} = \theta^*_i + \epsilon^*_{ik,s} \quad \epsilon^*_{ik,s} \sim \mathcal{N}(0, (\sigma^*_i)^2), \\ &\sigma^*_i = \exp(\eta^*_{\sigma_i}), \end{aligned} \tag{34}$$

for animal $k = 1, \ldots, n_i$ in group $i =$ C,E of the simulated data set $s = 1, \ldots, 10000$. Hence, scenarios were examined where the SD in the treatment group differed from the SD in the experimental group by a factor of $r_\sigma = \sigma^*_E / \sigma^*_C = \exp(\eta^*_{\sigma_E} - \eta^*_{\sigma_C}) \neq 1$ and various degrees of heteroscedasticity were examined in terms of different values of r_σ. The data in (34) were fitted with a Bayesian distributional model, where linear predictor terms are specified not only for the mean but also for the residual SD, allowing also the intercept $\eta^*_{\sigma_i}$ of the logarithmic residual SD to vary by group $i = C, E$ [1, 78]. The SDs were modeled by log-normal distributions [1]. This means that the logarithms of the SDs were modeled by normal distributions

$$\eta^*_{\sigma_i} \sim \mathcal{N}(m_{\sigma_i}, s^2_{\sigma_i}) \quad i = C, E.$$

The corresponding prior distributions need the specification of two parameters, the mean m_{σ_i} and the SD s_{σ_i} for the logarithmic SDs in each group (unlike the homoscedastic case where a common distribution of the SD was specified for both groups). Corresponding to the homoscedastic case, both mean parameters m_{σ_C} and m_{σ_E} were set to the expected value, that was obtained from a maximum likelihood (ML) fit of a normal distribution to the logarithm of the predicted SD from the historical data. Hence, the expected value of both SD prior distributions was the same as in the homoscedastic case. In contrast, for the SD parameters s_{σ_C} and s_{σ_E} different values were used for the experimental and the control group with the aim to obtain a less informative prior in the experimental group. More specifically, the SD parameter s_{σ_C} was set to the value, that was obtained in the ML approximation of the logarithm of the predicted SD from the historical data. The SD parameter s_{σ_E} instead was set, according to the concept of an UIP, to a value that results in σ_E having the SD of a single typical animal in the historical data. The intention behind this choice of parameters is that the prior for σ_E represents weaker information than the prior for σ_C. The reasoning for this choice of prior is the same as the reasoning for the parameter choice for the mean parameter, which was outlined in Sections 2.2.3 and Ṫhe different options and their relation to the historical data are summarized in Table 15.

Table 15: Different options for the prior distributions for the mean in the control group (MeanC) and for the standard deviation (SD) in the new experiment, whose effects on the power were compared in sensitivity analyses, and their relation to the historical data. ML: maximum likelihood estimate. UIP: unit information prior with residual standard deviation corresponding to a typical observation. In the heteroscedastic case, the residual standard deviation was allowed to vary by group with a less informative SD in the control group ($s_{\sigma_E} > s_{\sigma_C}$). For the experimental group mean, a UIP was used in each set-up.

Variable	Prior	Distribution	Reference in the historical data
MeanC	Predictive	$\mathcal{N}(m_C, s_C^2)$	ML estimates from the predicted mean
MeanC	UIP	$\mathcal{N}(m_C, \hat{\sigma}_C^2)$	m_C: ML estimate from the predicted mean, $\hat{\sigma}_C$: Median of the predicted SD
SD	Predictive	Log-normal(m_σ, s_σ^2)	ML estimates from the predicted SD
SD	None	Point-estimate $\hat{\sigma}_C$	Median of the predicted SD
SD	Heteroscedastic	Log-normal$(m_{\sigma_i}, s_{\sigma_i}^2)$, $i = C, E, s_{\sigma_E} > s_{\sigma_C}$	ML estimates from the predicted SD

2.5 Fictive multi-experiment example

For the C5aR1 application example, that was described in the previous section, the results from a Bayesian design analysis shall be compared to those from a frequentist analysis to see if and how much the sample size of the control group in the new experiment can be reduced in the Bayesian analysis by incorporating historical data as prior information. For both experimental outcomes in the C5aR1 example, only internal data went into the Bayesian model of the historical information. To investigate if there is an advantage of extending models of historical data with single experiments to combine multiple experiments, additional data were generated in a fictive meta-analysis example, as introduced in Example 2. Fitting a meta-analysis model and using it for the prediction of the new control allows for relaxing the strict assumption of identical parameters in the historical and the new experiments by modeling heterogeneity in the group means. The sample size results with a Bayesian predictive approach will be compared to the criticized frequentist sample size calculation approach from Example 1.

2.5.1 Bayesian model of the historical and the new data

The simulated data consisted of five historical experiments. The historical experiments were meant to represent a collection of translational animal experiments that address the same pathological phenotype. Each experiment was assumed to include characteristic measurements for the pathological phenotype from a control and an experimental group. The control groups were

thought to represent the pathological phenotype in the same mouse model, while in the experimental groups, different treatment candidates were thought to be investigated. The data were modeled as

$$\begin{aligned} y_{igk} &= \mu_i + \gamma_{ig} + \epsilon_{igk}, \\ \epsilon_{igk} &\sim \mathcal{N}(0, \sigma^2), \\ \gamma_{ig} &\sim \mathcal{N}(0, \tau_i^2) \end{aligned} \tag{35}$$

for animal $k = 1, \ldots, n_g$ in experiment $g = 1, \ldots, 5$ of group $i = C, E$ (control, experimental). The average difference between the measurements in experimental and control animals was chosen to be $\mu_E - \mu_C = 1.5$ and is again denoted by δ. The choice $\delta = 1.5$ represents a very large treatment effect according to the categorization by Cohen [148] and with respect to a within-group standard deviation (as SD of the animals within the treatment groups of one experiment) of $\sigma = 1$. The variance terms γ_{ig} introduce heterogeneity in the treatment means (as between-experiment SD or SD in animals from the same treatment group but from different experiments) that is due to the interaction of the experiments $g = 1, \ldots, 5$ and the treatments $i = C, E$ [25]. The historical control data were modeled to vary around the overall control mean with (according to the categorization in Table 5) moderate heterogeneity $\tau_C = 0.125$, while the historical data in the experimental groups were modeled to vary around the overall treatment mean with large heterogeneity $\tau_E = 0.5$. Modeling a larger heterogeneity of the experimental means than the heterogeneity of the control means reflects the assumption that the experimental groups were assumed to represent different treatment candidates that are only expected to be weakly related (e.g., the knockout of genes that are connected to the same biological pathway or that regulate the functionality of the same protein). In contrast, the control groups were assumed to represent the same pathological mouse model, and variation there is only thought to occur due to minor differences in the experiments, like different animal batches or housing conditions. The historical experiments were simulated with balanced designs (as equal sample sizes in the treatment and control group of each experiment) with typical sample sizes in animal experiments of ten and below [8].

The simulated historical data were assumed to be related to a newly planned animal experiment by sharing an overall mean and having identical standard deviations:

$$\begin{aligned} y_{i*k} &\sim \mathcal{N}(\theta_{i*}, \sigma_{i*}^2), \\ \theta_{i*} &\sim \mathcal{N}(\mu_i, \tau_i^2), \\ \sigma_{i*} &= \sigma_C, \end{aligned} \tag{36}$$

where $*$ indexes the new experiment with groups $i = C, E$ (control, experimental) and μ_i, τ_i and σ are the overall mean, the heterogeneity and the standard deviation of the historical data, as

defined above. Hence, the experiment-specific mean parameters θ_{ig} $g = 1, \ldots, 5, *$, $i = C, E$ in the historical and the new data of each group were modeled to vary around a common mean μ_i, according to the assumption (e) in Table 7.

Following the approach from Figure 17, a Bayesian (meta-analysis) model was fit to the historical control data, with prior distributions chosen according to the equations (14) and (15). The posterior distributions of the meta-analysis model were used to make predictions for the new experiment under the model (e) from Table 7 corresponding to the true model (36). As discussed in Chapter 4, the meta-analysis did not include the experimental data, reflecting the common situation that no particular data is available on the new treatment candidate, and the data from available experimental groups with different experimental treatments are assumed to be less similar than the control data [8]. Based on the predicted standard deviation from the historical control data and a minimal biologically relevant effect size (specified by the researcher as discussed in Section 4.1.2), a new frequentist sample size was calculated. The resulting sample size was compared to the results in the commonly used approach from Example 1, where the sample size is calculated based on the estimates from one single historical experiment and all parameters in the historical and the new experiments are assumed to be identical. In this common approach, the sample size was calculated with the frequentist Welch test for unequal variances and not the t-test, since the empirical variances in the control and experimental treatment group differed. As frequentist estimates, the empirical SDs from the historical experiments and the difference in the arithmetic means of the treatment groups were used.

The calculated sample sizes with either the estimates from the single historical experiments or the meta-analytic estimates from the combined control experiments were compared to the sample size that would result in data simulated with the true overall parameter values of the data (36) (instead of the estimated predictions from the meta-analysis or the experiment-wise estimates). Then, it was checked if the number of control animals that would be required with a frequentist analysis of the new data could be reduced by including the predictive distributions for the control mean and SD as prior distributions in a Bayesian analysis. This was done by following the simulation-based design analysis approach from Section 2.3. The predictions were used as prior distribution for the mean and SD of the new control group. The prior distributions for the new experimental group were set independently of the historical data, according to the considerations in Section 2.2.3. Alternatively, prior distributions for the mean and SD in the experimental group could have been obtained from the historical experimental data. However, this would result in a non-zero mean of the prior distribution for $\delta = \theta_E - \theta_C$, which is generally not recommended [146].

2.5.2 Simulation of prior-data conflicts

The simulation approach involves the repeated generation of hypothetical control and experimental data for the new experiment. This hypothetical data could be generated from the true parameters, which were used for simulating the historical data. However, this would not reflect the real-world situation of experimental planning in animal experiments, where the true mean difference and SDs are unknown. Hence, instead of simulating the hypothetical data from the true parameters, the data were simulated from the predictions for the mean and SD, which were estimated from the meta-analysis model of the historical data. Then it was investigated how much the power and type I error rate, which resulted from the final sample size decision, would change if the data in the design simulation were generated from the true parameters instead of the historical estimates. This mimics a type of "prior-data" conflict [38, 61–64] as the true mean parameters in the new data deviating from the expected value of their prior distributions.

In general, the notion of a prior-data conflict refers to a situation, where there is a disagreement between the information on the likelihood distribution parameters obtained from the observed data and the information on the parameters coming from the chosen prior distributions. In particular, this holds true in the situation when those parameter values, that are supported by the empirical estimates from the observed data, lie in the tail of their prior probability distributions [61]. In addition to deviations of the expected values of the prior distributions for the means from the true mean parameters, deviations of the expected value of the SD were investigated. Finally, the simultaneous deviation of the true means and residual SDs from the mean and median values of the prior distributions were investigated.

2.6 Computation in R

All analyses were performed with the statistical software `R` [66] and with software packages and estimation methods that were presented in my previous contribution, [1]. For MCMC computation, the `Stan` software [149] and the interface `RStan` [150] were used. The software uses MC algorithms like Gibbs-Sampling [151], as a special case of Metropolis-Hastings updates [97], and its improvements of Hamilton Monte-Carlo [152] and the No-U-Turn sampler [153]. For fitting the Bayesian models to the historical and the simulated new data the `brms` package [65] was used with its default MCMC parameters of four chains and 2000 iterations per chain, including 1000 warm-up iterations and a thinning factor (deciding the rate of iterations that are kept for analysis) of one. To decide whether the MCMC draws were likely to have converged against a target posterior distribution, characteristics of the MCMC chains were examined in convergence

diagnostic [47]. More specifically, MCMC trace plots, autocorrelation plots, the effective sample size, and the $\hat{R}$ statistic were examined [59]. The plausibility of different candidates of prior distributions in the Bayesian model of the historical data and the predictions for the new model were graphically investigated in prior predictive checks as introduced in Section 2.2.3, in my previous contribution [1] and in Example 6 using random number generating functions in `R` [66]. The aim of the choice of prior distributions for the historical data model was not to choose prior distributions that exactly reflect the distribution of the historical data, but rather to choose prior distributions that let the data have the main impact on the posterior distributions while defining an a priori range of plausible parameter values. In contrast, the aim of the choice of prior distributions for the new experiment's control data was to reflect the historical information in informative prior distributions. These informative prior distributions may have a major impact on the posterior distributions. The MCMC representation of the predictive distributions for the mean and SD in the new animal experiments were approximated by parametric distributions using the `fitdistrplus` package [154] and the `automixfit` function of the `RBesT` package [106]. The `automixfit` function uses the expectation-maximization (EM) algorithm to fit a series of mixture distributions with an increasing number of mixture components and selects the best fit according to the AIC value (which penalizes model complexity).

For the simulation-based design analysis from Section 2.3 hypothetical data were simulated with random number generating functions in `R` and for each simulated data set a Bayes factor was calculated with the `hypothesis` function of the `brms` package [65] and Bayesian equal-tailed credible intervals were calculated as sample 2.5% and 97.5% quantiles (as defined in (19)) of the difference of the posterior draws $\delta^*_{sr} = \theta^*_{E,sr} - \theta^*_{C,sr}$ for the posterior draws $r = 1, \ldots, R$ in the simulated data set $s = 1, \ldots, S$ of the mean parameters in the two groups. For the Bayesian models that were fit in the simulation-based design analysis, the number of MCMC iterations was set to $R = 10000$, including 2000 warm-up iterations, as suggested for the use of Bayes factors [146]. The power of the two-sample t-test and Welch test were calculated with the `power_t_test()` function from the `MESS` package [76]. The test's degrees of freedom were chosen depending on whether the design was simulated under the assumption of homoscedasticity (t-test) or heteroscedasticity (Welch test). Frequentist point estimates and confidence intervals were calculated with the `lm` function in `R` [66]. Bayesian equal-tailed credible intervals were compared to the highest density credible intervals estimated with the `bayestestR` package [155]. To account for heteroscedasticity in the Bayesian analysis, the residual SD was allowed to vary by group by using a distributional model in the formula of the `brms` function [65, 78]. In the frequentist model, heteroscedasticity was accounted for by the estimation of robust confidence intervals with the `sandwich` package [156], which estimates heteroscedasticity consistent (HC) variances. A HC estimator of type 3 was chosen, which is recommended for small sample sizes [157].

3 Results

In the following chapter, the theoretical considerations on sample size determination with Bayesian predictive approaches are illustrated in the C5aR1 experiment with the two experimental outcomes "bending stiffness" and "relative bone volume", as introduced in Section 2.4 and in my previous contribution [1]. The first part of the chapter consists of two subsections, corresponding to the two experimental outcomes. The two subsections are subdivided as follows: first, the historical data are analyzed descriptively. Second, the results from fitting a Bayesian model to the historical data are presented and the predictive distributions for the control group mean and the residual standard deviation (SD) are approximated by parametric distributions. Third, the results from a Bayesian sample size calculation are presented and the power of the chosen sample size is compared to the power with alternative prior distributions. Furthermore, additional design quality criteria are analyzed in the resulting design candidates, such as the root mean square error (RMSE) of point estimates of the posterior treatment mean difference, the exaggeration factor, as ratio by which significant estimates are expected to be exaggerated compared to the true parameter value, and properties of the estimated credible intervals, as Bayesian versions of frequentist confidence intervals. Finally, the sensitivity of the calculated sample size to the assumption of homoscedasticity (equal variances in both treatment groups) is examined. The sample size results are compared to the results from a frequentist sample size calculation with the two-sample t-test for homoscedastic designs and a Welch test for heteroscedastic designs and the additional design criteria in the Bayesian models are compared to the corresponding frequentist regression models.

In the second part of the chapter, the fictive example of historical data from multiple small experiments, which was used to illustrate the problems in common frequentist sample size calculation, is reanalyzed with Bayesian meta-analytic predictive approaches. The results complement those from my previous contribution [1] by providing additional applications of designing animal experiments by Bayesian predictive approaches while illustrating the proposed systematic procedure for sample size planning and illustrating both the benefits of using predictive prior distributions in Bayesian analysis and the benefits of using meta-analytic methods in the historical control data. Throughout this chapter, the main results are highlighted in red boxes.

3.1 Results for the outcome bending stiffness

3.1.1 Descriptive analysis of the historical data

The historical data for the outcome bending stiffness, which were provided by the researcher of the new animal experiment, consist of 48 female mice with a C57BL background from two different breeding laboratories (National Institute of Health (C57BL/6N) and the Jackson Laboratory (C57BL/6J)) and three different operation groups (no operation, sham operation, and ovariectomy). The data are presented in Table 16. The historical control animals, that are of interest for the new experiment, are those from the ovariectomized group.

Table 16: Group size (n), arithmetic mean, and empirical standard deviation (SD) of the extrinsic bending stiffness in the historical data that was provided by the experiment's researcher. The values are grouped by the factors operation (none, ovariectomy (Ovx) and Sham), age, and background strain and are represented on the square root scale (in Newton · millimeter).

Operation	Age	Background strain	n	Mean	SD
None	12	C57BL/6J	6	46.0	4.9
Ovx	16	C57BL/6J	7	18.5	3.2
Ovx	20	C57BL/6J	7	21.3	5.6
Ovx	52	C57BL/6N	7	5.6	3.9
Sham	16	C57BL/6J	7	28.5	3.5
Sham	20	C57BL/6J	6	31.6	4.1
Sham	52	C57BL/6N	8	29.7	4.3

3.1.2 Results from the Bayesian model of the historical control data

A Bayesian model is fit to the measurements of the bending stiffness in the historical animals from the ovariectomized group (as relevant control group for the new experiment). Predictions for the mean and the SD of further ovariectomized animals in the new experiment are made under the assumption that the new experiment represents additional data to the historical experiment (model (a) in Table 3).

Posterior draws

Table 17 presents the posterior MCMC draws of the model parameters from the Bayesian model fit to the internal historical data. The analyzed MCMC convergence diagnostics (as described

in Section 2.6) showed no signs of divergences. On average, the estimated posterior mean in Table 17 is slightly higher than the arithmetic mean of the bending stiffness in the observed ovariectomized animals from Table 16 (that correspond to the intercept's age group of 20 weeks). This is because the prior distribution for the intercept in the Bayesian model was centered at the estimate of the historical animals without operation (as independent information outlined in Section 2.4.2), which was higher than in the ovariectomized animals.

Table 17: Summary statistics (arithmetic mean, empirical standard deviation (SD), median and 95% equal-tailed interval) of the posterior draws of the model parameters (intercept μ, residual SD σ, age effects of the younger and older animals of age 16 and 52 weeks vs. the middle age of 20 weeks) from a Bayesian model fit to the researcher's historical data for bending stiffness in the ovariectomized mice on the square root scale using Markov Chain Monte Carlo methods.

Parameter	Mean	SD	Median	95% Interval
μ	22.90	1.94	22.80	(19.4,27)
β_{Age16}	-2.80	2.66	-2.80	(-8,2.6)
β_{Age52}	-15.70	2.65	-15.70	(-21,-10.4)
σ	4.90	0.95	4.70	(3.4,7.1)

Approximation of the MCMC draws and definition of prior predictive distributions

In Table 18 the maximum likelihood (ML) parameter estimates are presented, that result from the parametric approximations of the MCMC predictive distributions for the mean and residual SD in the new experiment's control group by normal and log-normal distributions. The resulting parametric probability density functions are presented together with the MCMC distributions in Figure 19.

Table 18: Maximum likelihood parameter estimates (Parameter 1, Parameter 2) that result from parametric approximations of the predicted θ_C and standard deviation σ_C for the square-root bending stiffness in the new control animals using the historical data from the ovariectomized animals. For the normal approximation of θ_C, the parameters are the mean and residual standard deviation (SD). For the log-normal approximation of σ_C, the parameters are the mean and SD of the logarithmic transformation of the predicted σ_C (meanlog, SDlog).

Variable	Distribution	Parameter 1	Parameter 2
θ_C	Normal	Mean=22.93	SD=1.94
σ_C	Log-normal	Meanlog=1.56	SDlog=0.19

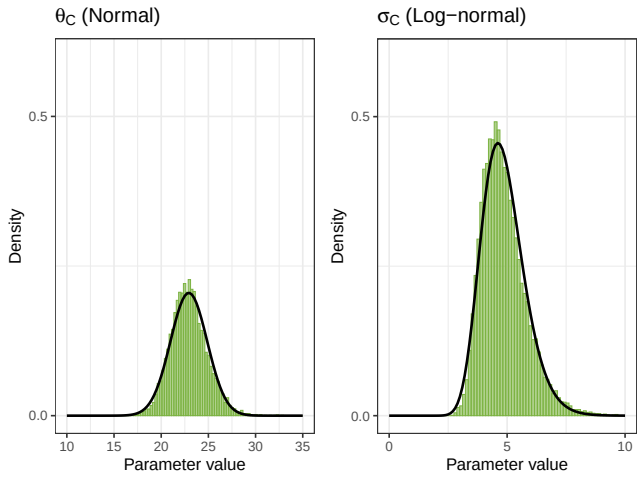


Figure 19: Parametric approximations of the predicted mean θ_C and standard deviation σ_C of the new control group's bending stiffness on the square root scale. Presented are histograms of Markov Chain Monte Carlo draws (green) and the probability density functions of the normal and log-normal distributions that were fitted via maximum likelihood estimation (black curves).

To facilitate the discussion of the predictions for animal researchers, hypothetical bending stiffness values for new control animals could be generated from the fitted parametric distributions, similar to the prior predictive checks from Example 6 and the application in my previous contribution [1]. To further facilitate the interpretation, the obtained bending stiffness values on the square root scale could then be back-transformed to the original scale by taking the square of the values.

3.1.3 Design analysis and sample size determination

The following section refers to the second part of the sample size planning procedure from Figure 17. It intends to compare the sample size from a classical calculation with the frequentist t-test to the sample size, that is necessary in a Bayesian analysis, where informative prior distributions (that are predicted from the historical data) are assigned both for the mean and the residual SD. The sample size calculation in both the frequentist and the Bayesian framework requires assumptions on the parameters in the new experiment. These assumptions are based on the Bayesian model that was fitted to the historical data, as described in Section 2.3 and Section 2.4.2.

Sample size determination

It is assumed that the researcher of the new experiment has specified a mean difference of $\delta^* = 6.5$ as minimal (biologically or clinically) relevant mean difference (which represents an expectation about the experimental outcome that is based on the researcher's experience or published literature, as further discussed in Section 4.1.2). The frequentist sample size calculation requires an estimate for the SD in the new experiment. This estimate is set to $\hat{\sigma} = 4.72$ according to the median of the prediction from the historical data. This would correspond to a sample size of $n^* = 10$ with the frequentist two-sample t-test, for which actually mean differences of at least $\delta = 6.26$ could be detected.

To investigate if and how much the necessary sample size in the control group could be reduced by incorporating the information from the historical data into the new experiment, the mean difference $\delta = 6.26$ is used to estimate the power, that results from a corresponding Bayesian analysis. The Bayesian models are fitted to simulated data for a series of decreasing sample sizes in the control group, starting from the value of $n_C = 10$ which was obtained from the frequentist sample size calculation. The sample size in the experimental group is kept at $n_E = 10$ since no historical data on the experimental treatment is available. The minimum control group sample size (with the restriction to be at least three animals per group), that is needed to detect the mean difference $\delta = 6.26$ in a Bayesian analysis with informative prior distributions, is presented in Table 19.

Results 1: Sample sizes for the bending stiffness example

Table 19 shows that, while with the frequentist t-test ten animals per group are necessary to achieve a power of at least 80%, in the Bayesian analysis (with either the Bayes factor and a threshold of three indicating moderate evidence according to the classification in Table 9 or the 95% equal-tailed credible interval not including the value zero) the sample size in the control group can be reduced to four animals, while still obtaining a power of over 80% and a type I error rate of only 1%.

Table 19: Sample size results (nC, nE) for the new experiment's control and experimental group for the outcome bending stiffness obtained with a standard deviation of 4.72 and a target power of at least 80% for testing the null hypothesis of equal group means $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$. The results from Bayesian models using the predictions from the historical data as prior information are compared to the results from a standard frequentist two-sample t-test. The Bayesian hypothesis tests reject the null hypothesis either if the Bayes factor (BF) exceeds a threshold of 3 (indicating at least moderate evidence for the alternative hypothesis [56]) or if the 95% posterior equal-tailed interval (ETI) for δ does not include the value zero. Given a sample size candidate, the actual power of a hypothesis test is calculated as its rejection rate in data sets simulated with the minimal relevant difference $\delta^* = 1.5$, and the type I error rate (FPR, false positive rate) is calculated in data sets simulated with the null value $\delta_0 = 0$. For the Bayesian models, the frequentist control group sample size is reduced to the minimal value ≥ 3 that maintains a power of $\geq 80\%$ and a FPR $< 5\%$.

Model	Test	nC	nE	Power	FPR
Bayes	BF	4	10	0.81	0.01
Bayes	ETI	4	10	0.82	0.01
Frequ.	t-test	10	10	0.80	0.05

Sensitivity analyses and investigation of additional design criteria

To reflect the dependence of the sample size results on different options for the prior distributions, the power, that could be obtained in the Bayesian analysis with informative prior distributions in a design with $n_C = 4$ and $n_E = 10$ animals for effects $\delta = 6.26$, is compared to the power under different prior distributions (labeled as in Table 15) in a sensitivity analysis. The results are shown in Table 20.

Table 20: Sensitivity analyses for the Bayesian model with mean difference $\delta = 6.26$ and with nC $= 4$ and nE $= 10$ under equal variances with residual standard deviation SD $= 4.72$ with different prior distributions for the mean in the control group (Prior Mean) and the residual SD (Prior SD). Predictive: prior with mean and SD fitted from the historical data. None: partly Bayesian analysis with the SD modeled as a fixed value without a prior distribution. The power resulting from the Bayes factor exceeding a threshold of 3 as hypothesis test (indicating at least moderate evidence for the alternative hypothesis [56]) is shown.

Model	Test	nC	nE	Prior Mean	Prior SD	Power
Bayes	BF	4	10	Predictive	Predictive	0.81
Bayes	BF	4	10	Predictive	None	0.82
Bayes	BF	4	10	UIP	Predictive	0.57

While with the informative prior for the mean and residual SD of the control group, a power of 81% was reached, the unit information prior (UIP) prior has only a power of about 57%, which is roughly the same as the power, that corresponds to a frequentist t-test with that design. The option "none" for the prior of the residual SD refers to a partly Bayesian analysis, where the residual SD is not modeled as a probability distribution, but a fixed value. The fixed value was set equal to the estimate, that was used for simulating the hypothetical new data. This option yields in this example a similar power to the Bayesian analysis with informative prior distributions for both the mean and the residual SD. In Table 21 additional design criteria, as described in Section 2.3.3, of the chosen Bayesian design are presented and compared to the results of evaluating these criteria in a frequentist regression analysis with the same design.

Table 21: Comparison of the chosen Bayesian designs for which a power of 80% was reached (with sample sizes of nC $= 4$ and nE $= 10$) regarding different design quality criteria for alternative prior distributions for the mean and standard deviation in the control group (Prior Mean, Prior SD). Estimate: average of the estimates of the true mean difference δ in the treatment groups. M ratio: average exaggeration factor by which the estimate exceeds a true difference of $\delta \neq 0$ in those cases where the null hypothesis was rejected. RMSE: root mean square error. CI width: average width of the credible/ confidence interval. CI coverage: percentage of credible intervals that include δ. The results are compared to those from frequentist regression models (Frequ.).

δ	Model	Prior Mean	Prior SD	Estimate	CI width	CI coverage	M ratio	RMSE
0.00	Frequ.	None	None	0.01	11.82	0.95		2.81
0.00	Bayes	Predictive	Predictive	-0.00	8.17	0.99		1.68
0.00	Bayes	Predictive	None	-0.00	8.11	0.98		1.67
0.00	Bayes	UIP	Predictive	0.01	10.11	0.97		2.34
6.26	Frequ.	None	None	6.27	11.82	0.95	1.30	2.81
6.26	Bayes	Predictive	Predictive	5.68	8.20	0.98	1.00	1.77
6.26	Bayes	Predictive	None	5.69	8.11	0.98	0.99	1.76
6.26	Bayes	UIP	Predictive	5.69	10.15	0.96	1.16	2.41

The Bayesian point estimates of the mean difference δ were calculated as the mean of the Bayesian posterior distributions. On average, the Bayesian point estimates underestimate the true values $\delta > 0$. However, point estimates provide only a limited view from a Bayesian perspective, that describes parameters by a whole probability distribution. Further information on the parameter estimation can be obtained by considering its precision, as the width of the credible intervals or MSE. Compared to the frequentist confidence intervals and the Bayesian credible intervals with a UIP for the control mean, the average width of the credible intervals with informative predictive prior for the control mean is reduced over the widths in the former analysis and the MSE, suggesting a higher precision of the Bayesian estimation. Also, the exaggeration factor is decreased in the latter designs, which indicates that the Bayesian design with informative prior distribution is less likely to entail over-optimistic hopes in treatment candidates. As further discussed in Section 4.1.2 and Section 4.2.2, the increased precision and decreased exaggeration factor may be

additional arguments for animal researchers to favor a Bayesian predictive approach for the new experiments, besides the reduction in sample size, that was presented in Table 19. The Type S error rate was zero for all designs with $\delta > 0$ and is thus not shown.

In Table 22 the power, type I error rate, and additional design quality properties are presented for Bayesian and frequentist analysis in the hypothetical situation where the residual SD in the experimental group deviates from the control group SD, but a priori equal SDs were assumed. This situation was realized by modeling the prior distributions for both residual SDs to be identical, but allowing different posterior distributions in a distributional model, as explained in Sections 2.4.2 and 2.6. The results from a completely Bayesian analysis, where the residual SDs in the two treatment groups are modeled as a probability distribution, are compared to alternative results from a partly Bayesian analysis, where the residual SDs are modeled to be known and fixed to the median of the predicted distribution, as described in Section 2.2.3 and Section 2.4.2. As a further reference, the results from typical frequentist analysis of heteroscedastic designs are shown with the Welch test for hypothesis testing and confidence intervals using sandwich estimation, as described in Section 2.6.

Table 22: Power and Type I error rate (P(reject H0)) and quality criteria for different heteroscedastic versions of the design with nC $= 4$ and nE $= 10$ for which a power of 80% was reached with a mean difference of $\delta = 6.26$ in the Bayesian analysis with predicted prior distributions. The Bayesian model (Bayes) is estimated as a distributional model and "CI" refers to the equal-tailed credible interval. For the frequentist model (Frequ.) the power and type I error rate with the Welch test are presented and "CI" refers to frequentist robust confidence intervals with the sandwich estimator. Estimate: average of the estimates (as the arithmetic mean of the data or posterior distribution) of the difference in treatment group means δ. CI width: average width of the credible/ confidence interval. CI coverage: percentage of credible intervals that include δ. M ratio: average exaggeration factor by which the estimate exceeds δ in those simulated data sets where the null hypothesis is rejected. RMSE: root mean square error.

δ	SD(C)	SD(E)	Model	P(reject H0)	Estimate	CI width	CI coverage	M ratio	RMSE
0.00	4.72	3.54	Bayes	0.00	-0.008	7.90	1.00		1.41
0.00	4.72	3.54	Frequ.	0.06	0.006	11.76	0.94		2.63
0.00	4.72	4.72	Bayes	0.01	-0.005	8.21	0.99		1.66
0.00	4.72	4.72	Frequ.	0.06	0.009	12.63	0.95		2.81
0.00	4.72	5.90	Bayes	0.02	-0.002	8.53	0.98		1.92
0.00	4.72	5.90	Frequ.	0.05	0.011	13.68	0.95		3.03
0.00	4.72	7.08	Bayes	0.04	0.001	8.87	0.96		2.18
0.00	4.72	7.08	Frequ.	0.05	0.014	14.89	0.96		3.27
6.26	4.72	3.54	Bayes	0.88	5.757	7.93	0.99	0.97	1.50
6.26	4.72	3.54	Frequ.	0.48	6.266	11.76	0.94	1.25	2.63
6.26	4.72	4.72	Bayes	0.80	5.667	8.24	0.98	1.00	1.77
6.26	4.72	4.72	Frequ.	0.46	6.269	12.63	0.95	1.30	2.81
6.26	4.72	5.90	Bayes	0.73	5.568	8.58	0.97	1.03	2.04
6.26	4.72	5.90	Frequ.	0.43	6.271	13.68	0.95	1.37	3.03
6.26	4.72	7.08	Bayes	0.67	5.462	8.92	0.95	1.06	2.33
6.26	4.72	7.08	Frequ.	0.39	6.274	14.89	0.96	1.45	3.27

Table 22 shows that, for the situation where both the residual SDs in the control and the experimental group were simulated to correspond to the SD estimate from the historical control data, the power and type I error for both the completely Bayesian analysis and the partly Bayesian analysis with fixed SD are about the same. In contrast, in the situation where the residual SD in the experimental group deviates from the control SD (and is in contrast with the prior or fixed value), the power of the completely Bayesian analysis also deviates from the power of the partly Bayesian analysis. This problematic is connected to the topic of prior-data conflicts, as further discussed in Section 4.1.2. The Type S error rate was zero for all designs with $\delta > 0$ and is thus not shown.

3.2 Results for the outcome relative bone volume

3.2.1 Descriptive analysis of the historical data

In this section, the results from fitting a Bayesian model to the internal historical bone volume to tissue volume (BV/TV) data, that were provided by the new experiment's researcher, are presented. The data represent a subset of the BV/TV data in my previous contribution and are summarized in Table 23.

Table 23: Group size (n), arithmetic mean, and empirical standard deviation (SD) of the relative bone volume (BV/TV) in the historical data that were provided by the experiment's researcher. The values are grouped by the factors operation (ovariectomy (Ovx) and Sham) and age and are presented on the logarithmic scale (as ratios). The table is based on my previous contribution [1].

Operation	Age	Background strain	n	Mean	SD
Ovx	(16,24]	C57BL/6	12	0.1	0.93
Sham	(16,24]	C57BL/6	10	1.4	0.36

3.2.2 Results from the Bayesian model of the historical control data

A Bayesian model is fit to the measurements of the relative bone volume in the historical animals from the ovariectomized group (as relevant control group for the new experiment). Predictions for the mean and the SD of further ovariectomized animals in the new experiment are made under the assumption that the new experiment represents additional data to the historical experiment (model (a) in Table 3).

Posterior draws

The posterior distributions for the model parameters (intercept μ and residual SD σ) in the Bayesian model of the internal historical data are presented in Table 24.

Table 24: Summary statistics (arithmetic mean, empirical standard deviation (SD), median and 95% equal-tailed interval (ETI)) of the posterior draws of the model parameters (intercept μ, residual SD σ) from a Bayesian model fit to the historical data of the logarithmic relative bone volume.

Variable	Mean	SD	Median (ETI)
μ	0.33	0.27	0.31 (-0.167,0.919)
σ	0.97	0.20	0.94 (0.652,1.435)

Approximation of the MCMC draws and definition of predictive distributions

In Figure 20 the log-normal and normal approximation of the MCMC draws of the predicted distributions for the residual SD and the mean (as posterior distributions of the residual SD and of the intercept from the historical data model) with parameters fitted via maximum likelihood (ML) estimation are presented. The ML parameter estimates of the approximated predictions for the new control group's mean and SD are presented in Table 25.

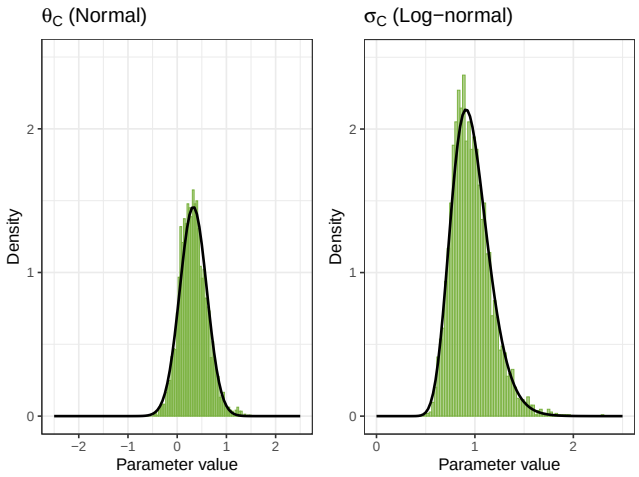


Figure 20: Parametric approximations of the predicted mean θ_C and standard deviation σ_C of the new control group's relative bone volume on the logarithmic scale. Presented are histograms of Markov Chain Monte Carlo draws and the probability density functions of the normal and log-normal distributions that were fitted via maximum likelihood estimation (black curves).

Table 25: Maximum likelihood parameter estimates (Parameter 1, Parameter 2) that result from parametric approximations of the predicted mean θ_C and standard deviation σ_C for the logarithmic relative bone volume in the new control animals using the historical data from the ovariectomized animals. For the normal approximation of θ_C the parameters are the mean and residual standard deviation (SD). For the log-normal approximation of σ_C the parameters are the mean and SD of the logarithmic transformation of the predicted σ_C (meanlog, SDlog).

Variable	Distribution	Parameter 1	Parameter 2
θ_C	Normal	Mean=0.33	SD=0.27
σ_C	Log-normal	Meanlog=-0.05	SDlog=0.2

The parametric approximations seem to provide again a tight fit of the posterior distributions. As stated for the outcome bending stiffness, hypothetical values of the relative bone volume values for new control animals can be generated from the fitted parametric distributions in prior predictive checks to facilitate the discussion of the predictions with animal researchers [1]. In addition, the data that is simulated in prior predictive checks could be back-transformed to the original measurement scale (by taking the exponent of the simulated values) to further facilitate their interpretation.

3.2.3 Design analysis and sample size determination

Sample size determination

It is assumed that the researcher of the new experiment has specified a mean difference of $\delta^* = 1.3$ as a minimal (biologically or clinically) relevant mean difference (as further discussed in Section 4.1.2). An estimate $\hat{\sigma}$ =0.94 for the residual SD is obtained from the Bayesian analysis of the historical control animals (as median of the prediction from model (a)). This would correspond to a sample size of $n^* = 10$ with the frequentist two-sample t-test, for which actually mean differences of at least $\delta = 1.25$ could be detected.

To investigate if and how much the necessary sample size in the control group could be reduced by incorporating the information from the historical data into the new experiment, the mean difference $\delta = 1.25$ is used to estimate the power that results from a corresponding Bayesian analysis. As for the outcome bending stiffness, the Bayesian models are fitted to simulated data for iteratively decreasing sample sizes in the control group, while keeping the sample size in the experimental group at the estimated value from the frequentist analysis of $n_E = 10$ (following the procedure from Figure 17). The minimum control group sample size (with the restriction to be at least three animals per group), that is needed to detect the mean difference $\delta = 1.25$ is presented in Table 26.

Results 2: Sample sizes for the micro-CT example

Table 26 shows that, while with the frequentist t-test ten animals per group are necessary to achieve a power of at least 80%, in the Bayesian analysis (with either the Bayes factor and threshold 3 indicating moderate evidence according to the classification in Table 9 or the 95% equal-tailed credible interval not including the value zero) the sample size in the control group can be lowered to three animals, while still obtaining a power of at least 80% and a type I error rate of only 1%.

Table 26: Sample size results (nC, nE) for the new experiment's control and experimental group for the outcome bending stiffness obtained with a standard deviation of 0.94 and a target power of at least 80% for testing the null hypothesis of equal group means $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$. The results from Bayesian models using the predictions from the historical data as prior information are compared to the results from a standard frequentist two-sample t-test. The Bayesian hypothesis tests reject the null hypothesis either if the Bayes factor (BF) exceeds a threshold of 3 (indicating at least moderate evidence for the alternative hypothesis [56]) or if the 95% posterior equal-tailed interval (ETI) for δ does not include the value zero. Given a sample size candidate, the actual power of a hypothesis test is calculated as its rejection rate in data sets simulated with the minimal relevant difference $\delta^* = 1.5$, and the type I error rate (FPR, false positive rate) is calculated in data sets simulated with the null value $\delta_0 = 0$. For the Bayesian models, the frequentist control group sample size is reduced to the minimal value ≥ 3 that maintains a power of $\geq 80\%$ and a FPR $< 5\%$.

Model	Test	nC	nE	Power	FPR
Bayes	BF	3	10	0.88	0.01
Bayes	ETI	3	10	0.89	0.01
Frequ.	t-test	10	10	0.80	0.05

Sensitivity analyses and investigation of additional design criteria

The power, that could be obtained in the Bayesian analysis with informative prior distributions in a design with $n_C = 3$ and $n_E = 10$ animals for effects $\delta = 1.25$, is compared to the power under different prior distributions (labeled as in Table 15). The results are meant to reflect the dependence of the sample size result on different options for the prior distributions shown and are shown in Table 27.

Table 27: Sensitivity analyses for the Bayesian model with true mean difference $\delta = 1.25$ and with nC$= 3$ and nE$= 10$ under equal variances with residual standard deviation SD $= 0.94$ with different prior distributions for the mean in the control group (Prior Mean) and the residual SD (Prior SD). Predictive: prior with mean and SD fitted from the historical data. None: partly Bayesian analysis with the SD modeled as fixed value without a prior distribution. The power resulting with the Bayes factor exceeding a threshold of 3 as hypothesis test (indicating at least moderate evidence for the alternative hypothesis [56]) is shown.

Model	Test	nC	nE	Prior Mean	Prior SD	Power
Bayes	BF	3	10	UIP	Predictive	0.52
Bayes	BF	3	10	Predictive	Predictive	0.88
Bayes	BF	3	10	Predictive	None	0.89

While with the informative prior for the mean and residual SD of the control group, a power of 88% was reached, the UIP prior has only a power of about 52%. With an informative prior distribution for the control mean, the partly Bayesian analysis (where only the man parameters are modeled by prior distributions, whereas the SDs are modeled as frequentist fixed values) yields a similar power to the completely Bayesian analysis with informative prior distributions for both the mean and residual SD in the control group.

In Table 28 the results from the additional design criteria from Section 2.3.3 are presented and compared to the results of these criteria in a frequentist regression analysis. The Type S error rate was zero for all designs with $\delta > 0$ and is thus not shown. Similar statements like for the outcome bending stiffness can be made (underestimation of the effect in the Bayesian analysis, especially with the predictive prior for the mean, but greater credible interval coverage with likewise smaller interval widths and lower MSE and smaller exaggeration ratios in the latter analysis).

Table 28: Comparison of the chosen Bayesian designs for which a power of 80% was reached (with sample sizes of nC=3 and nE=10) regarding different design quality criteria for alternative prior distributions for the mean and standard deviation in the control group (Prior Mean, Prior SD). Estimate: average estimated mean difference δ in the treatment groups. M ratio: average exaggeration factor by which the estimate exceeds a true difference of $\delta \neq 0$ in those cases where the null hypothesis was rejected. RMSE: root mean square error. CI width: average width of the credible/ confidence interval. CI coverage: percentage of credible intervals that include δ. The results are compared to those from frequentist regression models (Frequ.).

δ	Model	Prior Mean	Prior SD	Estimate	CI width	CI coverage	M ratio	RMSE
0.00	Frequ.	None	None	0.00	2.65	0.95		0.62
0.00	Bayes	Predictive	Predictive	-0.02	1.48	0.99		0.30
0.00	Bayes	Predictive	None	-0.02	1.47	0.99		0.30
0.00	Bayes	UIP	Predictive	-0.00	2.18	0.97		0.49
1.25	Frequ.	None	None	1.25	2.65	0.95	1.40	0.62
1.25	Bayes	Predictive	Predictive	1.12	1.49	0.98	0.95	0.32
1.25	Bayes	Predictive	None	1.12	1.47	0.98	0.94	0.32
1.25	Bayes	UIP	Predictive	1.13	2.18	0.97	1.21	0.51

Table 29 shows the power, type I error rate, and additional design quality properties for Bayesian and frequentist analysis in the situation where the SD in the experimental group deviates from the control group SD, but a priori equal SDs were assumed. The Type S error rate was again almost zero for all designs with $\delta > 0$ and is thus not shown. The a priori equal SDs are modeled by setting identical prior distributions for both residual SDs, while allowing different posterior distributions in a distributional model, as explained in Sections 2.4.2 and 2.6. The results are similar to those from the bending stiffness outcome and are discussed in Section 4.2.

Table 29: Power and Type I error rate (P(reject H0)) and quality criteria for different heteroscedastic versions of the design with nC $= 3$ and nE $= 10$ for which a power of 80% was reached with a mean difference of $\delta = 1.25$ in the Bayesian analysis with predicted prior distributions. The Bayesian model (Bayes) is estimated as a distributional model and "CI" refers to the equal-tailed credible interval. For the frequentist model (Frequ.) the power and type I error rate with the Welch test are presented and "CI" refers to frequentist robust confidence intervals with the sandwich estimator. Estimate: average of the estimates (as arithmetic mean of the data or posterior distribution) of the difference in treatment group means δ. CI width: average width of the credible/ confidence interval. CI coverage: percentage of credible intervals that include δ. M ratio: average exaggeration factor by which the estimate exceeds δ in those simulated data sets where the null hypothesis is rejected. RMSE: root mean square error.

δ	SD(C)	SD(E)	Model	P(reject H0)	Estimate	CI width	CI coverage	M ratio	RMSE
0.00	0.94	0.70	Bayes	0.00	-0.019	1.41	1.00		0.24
0.00	0.94	0.70	Frequ.	0.07	-0.000	2.62	0.92		0.59
0.00	0.94	0.94	Bayes	0.01	-0.018	1.49	0.99		0.29
0.00	0.94	0.94	Frequ.	0.07	0.000	2.78	0.94		0.62
0.00	0.94	1.17	Bayes	0.02	-0.017	1.57	0.97		0.35
0.00	0.94	1.17	Frequ.	0.06	0.001	2.97	0.95		0.66
0.00	0.94	1.41	Bayes	0.05	-0.016	1.64	0.95		0.41
0.00	0.94	1.41	Frequ.	0.06	0.001	3.20	0.95		0.71
1.25	0.94	0.70	Bayes	0.95	1.135	1.42	0.99	0.93	0.26
1.25	0.94	0.70	Frequ.	0.36	1.250	2.62	0.92	1.28	0.59
1.25	0.94	0.94	Bayes	0.88	1.115	1.50	0.98	0.95	0.32
1.25	0.94	0.94	Frequ.	0.36	1.250	2.78	0.94	1.33	0.62
1.25	0.94	1.17	Bayes	0.79	1.094	1.58	0.96	0.98	0.38
1.25	0.94	1.17	Frequ.	0.35	1.251	2.97	0.95	1.39	0.66
1.25	0.94	1.41	Bayes	0.71	1.071	1.66	0.94	1.01	0.45
1.25	0.94	1.41	Frequ.	0.32	1.251	3.20	0.95	1.48	0.71

3.3 Total sample size for the real-world example

If a probability of at least 80% of the Bayesian hypothesis test to detect a true mean difference of at least the specified minimally relevant value (while limiting the probability of false rejections of the null hypothesis of equal group means by 5%) for only the outcome relative bone volume is considered a sufficient indication for an improvement in the bone quality, then the lower sample size of three animals in the control group can be considered as group sample size for calculating the total sample size in the C5aR1 experiment. In contrast, if the researcher wants to be able to detect the specified minimally relevant effects for both experimental outcomes, the larger sample size of four animals in the control group should be chosen. The latter scenario leads, with the two specific knockouts of the C5aR1 in osteoblasts and osteoclasts from Section 2.4 as experimental groups and their corresponding control groups without the knockout, with the Bayesian predictive approach to a total sample size of

$$N_{\text{Bayes}} = 2 \cdot (n_{C,\text{Bayes}} + n_E) = 2 \cdot (4 + 10) = 28,$$

whereas the former scenario would lead to a total sample size of $N_{\text{Bayes}} = 2 \cdot (3 + 10) = 26$. In comparison, the total sample size under the frequentist approach would be:

$$N_{\text{Frequ.}} = 2 \cdot (n_{C,\text{Frequ.}} + n_E) = 2 \cdot (10 + 10) = 40.$$

Hence, the total sample size reduction in the Bayesian over the frequentist approach would be $40 - 28 = 12$ animals or $40 - 26 = 14$ animals. If additional sham animals are included in the experiment, then the total sample size benefit from the Bayesian approach may yet be larger and the exact number depends on whether historical information is included for the sham groups, as further discussed in Section 4.1.1. For the experimental group, the sample size in the Bayesian approach was not reduced over the number obtained in the frequentist approach, since no historical data on the experimental treatment were provided. As a consequence, it is beneficial in the Bayesian approach to assign a greater proportion of the total sample size to the experimental group than to the control group, whereas, in the frequentist approach, an equal allocation ratio is optimal (in the case of homoscedasticity), as further discussed in Chapter 4.

3.4 Results for the fictive multi-experiment example

In this section, the results from a classical frequentist sample size calculation based on a single small historical experiment are compared to those obtained in a Bayesian predictive approach using a meta-analysis model of the historical control data. Furthermore, the sensitivity of the sample size results to prior-data conflicts, in terms of the expected values of the prior distributions deviating from the true parameters of the data, are examined.

3.4.1 Descriptive analysis of the historical data

The historical data consists of five small experiments, as hypothetical measurements of an experimental outcome. The control data were introduced in Example 2 and are simulated with only moderate heterogeneity. The control data are supplemented with experimental data representing different experimental treatments with large heterogeneity in the treatment means, as introduced in Section 2.5. A summary of the data is presented in Table 30. The greater heterogeneity is reflected in the, on average, greater deviations of the experiment-wise means from their overall mean in the experimental groups than in the control groups. More precise statements about the model parameters can be made by fitting a meta-analysis model to the data, as presented in the next section.

Table 30: Group size (n), arithmetic mean, and empirical standard deviation (SD) from five hypothetical historical experiments. The data represent a collection of historical experiments addressing the same pathological phenotype as in the new experiment, where each experiment compares fairly similar control animals (C) to animals with different experimental treatments (E).

Experiment	Group	n	Mean	SD
1	C	5	0.20	0.90
1	E	5	1.89	1.06
2	C	5	-0.40	1.04
2	E	5	1.64	1.25
3	C	7	-0.08	1.13
3	E	7	0.85	0.98
4	C	10	0.51	0.63
4	E	10	0.68	0.93
5	C	10	0.08	1.12
5	E	10	1.24	0.90

3.4.2 Results from the Bayesian model of the historical control data

In the first step of the proposed sample size planning procedure, a Bayesian individual animal data meta-analysis (IADMA) model is fit to the historical control data with the aim to obtain predictions for the mean and the SD of the new experiment which can be used as prior distributions and planning basis for the new experiment. Only the control data (plus a biologically minimal relevant effect size) are used for planning the new experiment. As discussed in Chapter 4, this is because, due to the quite heterogeneous historical and new experimental treatments and the lack of data for the particular new treatment, the historical experimental are assumed to not provide a good information source for the experimental treatment in the new experiment. Alternative scenarios are discussed in Section 4.1. In Table 31, a summary of the posterior parameter distributions from the Bayesian meta-analysis model of the historical control data is presented. On average, the overall control mean is slightly overestimated with this historical data as compared to the true mean $\mu_C = 0$. Also, the true heterogeneity $\tau_C = 0.1$ in the control group is slightly overestimated on average, whereas the true SD $\sigma = 1$ is estimated quite well on average. For comparison, an equivalent frequentist IADMA model was considered, but the model failed to converge due to the small sample sizes in the groups.

Table 31: Summary statistics (arithmetic mean (Mean), empirical standard deviation (SD), median and 95% equal-tailed credible interval (ETI)) of the posterior draws of the Bayesian meta-analysis model fit to the fictive historical multi-experiment control data.

Parameter	Mean	SD	Median	95% ETI
μ_C	0.13	0.21	0.12	(-0.27,0.56)
τ_C	0.24	0.19	0.19	(0.01,0.68)
σ	1.00	0.12	0.98	(0.79,1.26)

In Table 32 the parameters of the ML approximations of the predictive distribution for the new mean θ_{C*} and the new residual SD σ_{C*} in the control group are presented. As described in Section 2.5, the mean was predicted under the assumption that the historical and the new means are similar but not identical (model (e) in Table 7), while the SD was predicted under the assumption that the historical and the new SDs are identical (corresponding to model (a) in Table 7 with parameters representing the SDs). As discussed in Section 4.1, the weaker assumption of similarity could be realized for the SDs as well, if doubts exist on the assumption of identical parameters. Since the overall mean μ_C was on average slightly overestimated in the meta-analysis model, also the mean of the predicted θ_{C*} is slightly higher than the true overall mean. The probability density functions of the parametric approximations were presented together with the approximated non-parametric distributions of the MCMC draws in Figure 9.

Table 32: Maximum likelihood parameter estimates (Parameter 1, Parameter 2) that result from parametric approximations of the predicted θ_{C*} and standard deviation σ_{C*} for the new control animals using the meta-analysis model of the fictive historical multi-experiment control data. For the normal approximation of θ_{C*} the parameters are the mean and residual standard deviation (SD). For the log-normal approximation of σ_{C*} the parameters are the mean and SD of the logarithmic transformation of the predicted σ_{C*} (meanlog, SDlog).

Variable	Distribution	Parameter 1	Parameter 2
θ_{C*}	Normal	Mean$= 0.13$	SD$= 0.37$
σ_{C*}	Log-normal	Meanlog$= -0.01$	SDlog$= 0.12$

3.4.3 Design analysis and sample size determination

Sample size determination

The predictions for the control mean and SD are used to determine the necessary sample size for the new experiment by following the design simulation and sample size calculation procedure from Section 2.3. It is assumed that the researcher considers a mean difference of $\delta^* = 1.5$ as a minimal biologically or clinically relevant effect. The results are presented in Table 33 and are compared to the sample sizes that result from classical calculations with the estimates from only one of the single small historical experiments.

Results 3: Sample sizes for the fictive meta-analysis example

In the first five rows of Table 33, the results from the common approach of calculating the sample size in the new experiment with the estimates from only one single historical experiment, and assuming that the parameters in both the historical control and the experimental group are the same as the parameters in the new experiment's control and experimental group, are shown. In accordance with the observations in Example 1, the calculated sample size from this approach varies greatly, depending on the historical experiment that is used as a planning basis. The calculated values are up to 344 animals per group (row 4) and 60 animals per group on average. In comparison, when calculating the sample size with the true overall means, the result would be only $n = 9$ animals per group with either the two-sample t-test or the Welch test (row 6/7).

Table 33: Sample size results (nC, nE) for the fictive new experiment's control and experimental group required to obtain a target power of at least 80% for testing the null hypothesis of equal group means $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$. Row 1-5: results from the common approach of calculating the sample size with the estimates from only one single historical experiment and assuming both the parameters in the historical control (C) and the historical experimental (E) group are identical to the new experiment's treatment groups. Row 6-7: results with the true overall means and standard deviation (SD) of the meta-analysis model of the historical and the new data. Row 8-10 (highlighted): results from combining the historical control data in a meta-analysis and calculating the sample size with the meta-analytic estimate of the SD and a minimal biologically relevant effect size of $\delta^* = 1.5$ either with a frequentist t-test (row 8) or with the Bayesian predictive approach including the predicted prior distributions for the control mean and SD in a Bayesian analysis of the new experiment (row 9-10). Row 11-12: results from using the estimated Bayesian predictive prior distributions from the observed control data, but simulating the new experiment's data from the true parameters of the meta-analysis model (instead of the estimated predictions). The results with a Bayesian hypothesis test that rejects H_0 if the Bayes factor (BF) exceeds a threshold of 3 (indicating at least moderate evidence for the alternative hypothesis [56]) are compared to those from a Bayesian hypothesis tests that rejects H_0 if the 95% posterior equal-tailed interval (ETI) for δ does not include the value zero. Given a sample size candidate, the actual power of a hypothesis test is calculated as its rejection rate in data sets simulated with the minimal relevant difference $\delta^* = 1.5$, and the type I error rate (FPR, false positive rate) is calculated in data sets simulated with the null value $\delta_0 = 0$. For the Bayesian models (including the historical control information), the frequentist control group sample size is reduced to the smallest value ≥ 3 that maintains a power of $\geq 80\%$ and a FPR $< 5\%$.

	Planning basis	New experiment						
	(groups)	Model	Parameters	Test	nC	nE	Power	FPR
1	Experiment 1 (C,E)	Frequ.	Estimates	Welch	7	7	0.83	0.05
2	Experiment 2 (C,E)	Frequ.	Estimates	Welch	7	7	0.85	0.05
3	Experiment 3 (C,E)	Frequ.	Estimates	Welch	22	22	0.81	0.05
4	Experiment 4 (C,E)	Frequ.	Estimates	Welch	344	344	0.80	0.05
5	Experiment 5 (C,E)	Frequ.	Estimates	Welch	14	14	0.80	0.05
6	Meta-analysis (C,E)	Frequ.	True values	Welch	9	9	0.84	0.05
7	Meta-analysis (C,E)	Frequ.	True values	t-test	9	9	0.85	0.05
8	Meta-analysis (C)	**Frequ.**	Estimates	t-test	**8**	**8**	0.81	0.05
9	Meta-analysis (C)	**Bayes**	Estimates	BF	**3**	**8**	0.89	0.01
10	Meta-analysis (C)	**Bayes**	Estimates	ETI	**3**	**8**	0.89	0.01
11	Meta-analysis (C)	Bayes	True values	BF	3	8	0.84	0.01
12	Meta-analysis (C)	Bayes	True values	ETI	3	8	0.84	0.01

In the first step, the sample size results from the criticized approach (row 1-5 of Table 33) are compared to the sample size that results from combining the historical control data in a meta-analysis and calculating the sample size with the meta-analytic estimate of the

SD from Table 31 and a minimal clinically or biologically relevant effect size (specified by the researcher) of $\delta^* = 1.5$. As stated before and further discussed in Chapter 4, it is suggested to only fit a meta-analysis to the historical control data, but not the experimental data, under the assumption, that the experimental treatment candidates investigated in the historical experiments are quite different from the new experiment's treatment candidate, and so are the experimental data, whereas the control data in the historical and new experiments are quite similar. The results are presented in the eighth row of Table 33. The sample size with the meta-analytic approach using the control data and a biologically relevant mean difference deviates by only one animal per group from the sample size that results with the true meta-analytic parameters (row 8 vs. 6/7). In contrast, the sample size that results from using the estimates from only one experiment at a time, and assuming both the parameters in the historical control and experimental group are identical to the parameters in the new experiment, deviates by up to 335 animals from the sample size obtained with the true meta-analytic parameters (row 4 vs. 6/7).

In the second step, the frequentist sample size result in row 8 is compared to the result with the proposed Bayesian predictive approach where the meta-analysis model of the control data is not only used to estimate the SD for the new experiment but also to increase the information content of the new experiment by including the predicted prior distributions for the control mean and SD from Table 32 in a Bayesian analysis of the new data. With the sample size planning procedure from Section 2.3 it is examined whether the estimated sample size of 8 animals per group with the frequentist t-test (row 8 in Table 33) can be reduced by including the prior information on the new experiment's control group in a Bayesian analysis with either the Bayes factor based hypothesis test (24) or the ETI based hypothesis test (21). The highlighted results (row 9/10 vs. 8) suggest that, by using the prior distributions from the historical data, the sample size in the new control group can be lowered to the (here defined) minimum of three animals with either Bayesian test, representing a reduction of five animals compared to the frequentist t-test sample size estimate. Compared to the calculated sample size with the true values and a frequentist test (row 6/7), the sample size with the Bayesian predictive approach represents a reduction of five animals, and compared to the highest sample size estimate from the single historical experiments it represents a reduction of up to 341 animals (row 4).

In the last two rows of Table 33 the sample size with the Bayesian predictive approach, using the predicted prior distributions from Table 32 together with the meta-analytic control estimates for the design simulation of the new experiment, is compared to the sample size

that would result from using the same estimated prior distributions, but simulating the new experiment's data from the true parameters of the meta-analysis model (instead of the estimated meta-analytic predictions). As further discussed in Chapter 4, this scenario reflects the situation that the prior distributions may bias the posterior conclusions in cases of conflicts between the data and the prior distributions since they are only estimates of the true parameters. The results in the last two rows of Table 32 indicate, that, with the chosen sample size in the Bayesian predictive approach, the Bayesian hypothesis tests would still be powerful enough (with a power of over 80%) and could maintain a type I error rate of below 5% when analyzed with the true parameter meta-analytic parameters.

Sensitivity to further prior-data conflicts regarding the standard deviation

To examine scenarios where the expected value of the predicted SD deviates more from the estimated SD in the new experiment than in the results from these historical data, the power, and type I error rate are evaluated in design simulations where the parameters of the predicted prior SD are modified to be less similar to the true meta-analytic parameter values. The prior for the mean parameters are kept at the predicted values and the data are simulated from the true meta-analytic parameter values. The results are presented in Table 34.

Table 34: Sensitivity of the Bayesian sample size result to prior-data conflicts in terms of the expected values of the prior distributions deviating from both the true means of the new experiment's data (MeanC, MeanE) and the true standard deviation (SD). Various degrees of deviation of the expected prior SD from the true SD are examined, as indicated by the ratio E(Prior SD)/SD. As before, the prior distribution for the control mean and the common SD are obtained as predictions from the historical data (Table 32), and the prior distribution in the experimental group is centered at the control mean, but is chosen to be less informative (assuming a smaller degree of similarity between the historical and the new experimental data than between the control data). The power and type I error rate (FPR, false positive rate) of the completely Bayesian analysis, where both the mean and the residual SD are modeled as the predicted probability distributions ("Predictive"), are compared to the power where the residual SD is not modeled as probability distribution, but as a fixed value instead (as the median of the prediction) ("None"). The hypothesis test is based on whether the Bayes factor exceeded a threshold of 3 (indicating at least moderate evidence for the alternative hypothesis [56]).

MeanC	MeanE	SD	Prior SD	E(Prior SD)/SD	nC	nE	Power	FPR
0.0	1.5	1.0	Predictive	0.5	3	8	0.93	0.16
0.0	1.5	1.0	None	0.5	3	8	0.94	0.26
0.0	1.5	1.0	Predictive	1.5	3	8	0.84	0.01
0.0	1.5	1.0	None	1.5	3	8	0.85	0.02

According to Table 34, the larger expected value of the residual SD prior distribution of 1.5 times the true SD does not impact the power and type I error much as compared to the power and type I error, that were obtained with a SD whose expected value is equal to the true value. In the cases, where the data were simulated with a true mean difference of 1.5, the smaller expected value of the SD prior distribution of 0.5 times the true SD leads to an increase in both the power to 93% for the model with the SD as a random variable (completely Bayesian analysis) and 94% for the model with the SD as a fixed value (partly Bayesian analysis). In the cases, where the data were simulated with a true mean difference of zero, the smaller expected value of the SD's prior distribution leads to an increase of the type I error to 16% in the completely Bayesian analysis, as compared to a type I error rate of 26% in the partly Bayesian analysis.

4 Discussion

In this chapter, the material and methods, that were used to analyze the real-world and fictive animal data from the previous chapter, are discussed and compared to alternative methods in related scientific literature. Moreover, the sample size results from the Bayesian predictive approach are discussed and assessed in comparison to the results from classical frequentist approaches. A conclusion is then drawn on the introductory research questions, assessing, in particular, whether and to what extent the Bayesian predictive approach has the potential to improve the design of translational animal experiments. The chapter ends with an outlook for future steps, that are necessary to integrate the proposed methods into the daily practice of planning and analyzing translational animal experiments.

4.1 Discussion of the materials and methods

4.1.1 Analysis of the historical control data

Availability and use of the historical data

The here suggested methodology was illustrated for the two main outcomes of a real-world translational animal experiment investigating the effect of the C5aR1 receptor knockout on bone quality in mice under ovariectomy (C5aR1 experiment). Additionally, the methodology was illustrated for a fictive experimental outcome, where hypothetical data were simulated for multiple related experiments. For both outcomes in the C5aR1 experiment, internal control data from a historical experiment were provided by the researcher, who was planning the experiment. To increase the scope of the historical data, additional external data from the Mouse Phenome Database (MPD) were considered. The MPD is headquartered by the Jackson Laboratory in the United States of America and offers a great variety of open-access quantitative data on phenotyping outcomes and disease-related traits in both naive mice and treated mice [58, 136]. The MPD website has a clear structure, that makes it easy to find the relevant data. Summaries of the experiments' characteristics and their experimental outcomes helped to gain a fast overview of the experiments.

The data come in a tidy, standardized format and can easily be accessed in a comma-separated values format. which may then be further examined in spreadsheet editors or statistical computation tools, like `Excel` [158] and `R` [66]. In addition to the data, the MPD provides descriptions of the variables in the data. The comparisons of measurements from different experiments are facilitated by the common Mammalian Phenotype Ontology [137] (as phenotype terms, definitions, and synonyms that can be used to describe mammalian phenotypes) developed by the Mouse Genome Informatics (MGI) (https://www.informatics.jax.org/). Additionally, also the experiments' protocols are provided. These protocols may be useful when it comes to judging the similarity of the historical and the new data. The judgment of the similarity is essential for setting up the statistical model of the historical data and for the prediction, as further discussed below. The experiments' protocols may contribute to this judgment, for example, by providing details on the measuring procedures or machines, which may cause variation in the data. With these characteristics, the MPD is an attractive way to obtain additional data on an experimental outcome. Besides the MPD, publications on `pubmed` (https://www.ncbi.nlm.nih.gov/) and data from the International Mouse Phenotyping Consortium (IMPC) [159] were searched for additional phenotypic biological open-access data on the experimental outcomes in the C5aR1 experiment, but no suitable data could be found.

What, in the end, prevented the inclusion of the MPD data in the prediction model for the C5aR1 experiment was, that for both C5aR1 outcomes only data of mice was available, whereas the control group of main interest in the C5aR1 experiment were ovariectomized mice. In general, animal experiments typically include several control groups that can be distinguished based on their purpose regarding the experiment's research aim [80]. The pathological phenotype of interest in the C5aR1 experiment was postmenopausal osteoporosis, represented by ovariectomy in the mice, and the research aim was to investigate the effects of a C5aR1 knockout in this phenotype. Consequently, the comparisons of main interest were those between ovariectomized mice with and without the C5aR1 knockout in osteoblasts and osteoclasts. The untreated ovariectomized (without the C5aR1 knockout) are thus the positive controls in the experiments. In positive controls, a change (that is causing e.g. a pathological phenotype, which is here ovariectomy) is initiated as compared to the condition in naive, healthy animals. The comparisons of such positive control animals with treated experimental animals are then usually of main interest for the research aim. Besides positive control groups, animal experiments usually comprise additional negative, sham, or vehicle control groups, which are of secondary interest to the research aim [80]. In negative control groups no change, as compared to the naive, healthy population is initiated, while in sham and vehicle controls the treatment of the positive controls is imitated without the actual active procedure or substance. These control groups serve as a reference for judging the effects, that are observed in the positive controls. In the C5aR1 experiment, the ovariectomized control

group is supplemented with a sham-operation group to differentiate the effects of the ovariectomy from unspecific operation effects like anesthesia, the incision, or the perception of the operation [160]. In this thesis, only the positive controls were considered, since they are of main interest to the research aim and thus most relevant for the sample size calculation of the new experiment. However, the same Bayesian predictive approach may also be applied to further control groups in a new experiment. The inclusion of additional historical knowledge for comparisons featuring those additional controls is expected to have the same benefit regarding the sample size reduction in the new experiment, as it was demonstrated for the positive controls. Furthermore, the additional wild type data can be used to get an idea of the range of plausible treatment effects, that are examined in the design analysis, as illustrated in my previous contribution [1] and further discussed in Section 4.1.2.

Walley et al. [33] suggest replacing control animals in the new experiment with animals from historical experiments, depending on their role in the new experiment. For control animals that are of primary interest in the planned experiment (positive controls), they suggest using the historical data as prior distribution for the new controls. Meanwhile, for control animals, that are not of primary interest in the new experiment (negative, sham, or vehicle controls), they suggest replacing the control group in the new experiment completely by the estimates from the historical data. Such a complete replacement would seem very desirable from animal welfare aspects as outlined by the Replacement, Reduction and Refinement (3R) principles [2]. However, not collecting any data in a certain control group of the new experiment at all, would also come with the risk of biased conclusions in the new experiment. Especially when the historical data were not collected by the researcher of the new experiment, unmeasured changes in the experimental conditions are likely to cause random noise in the new data. This random noise is then not reflected in comparisons of the new data of treated animals with respective untreated historical control data. As a consequence, it remains unclear whether the effects, that are observed in those comparisons, are caused by the treatment or by the random noise.

A further use of control data from additional groups may be for the definition of plausible ranges of treatment effects. For example, the researcher may assume that the maximal effect of the treatment is to completely cure the pathological condition in the positive control. This would, for example, mean that (on average) the measurements in the new experiment's experimental group should equal the values in the negative control group. In this situation, the treatment effect of interest in the new experiment could be approximated by the difference in the means of negative historical controls (representing the condition of a complete cure of the pathological phenotype) and positive controls. The topic of setting a treatment effect for planning the new experiment is further discussed below. Besides the definition of prior distributions and treatment effects for

the new experiments, additional control data can help quantify heterogeneity in the outcome in different strains or laboratories. This could give an impression of how representative the internal data is for the mouse population in general concerning the experimental outcome. For example, if the measurements in internal wild type animals deviate greatly from external wild type animals, then additional care might be necessary for the interpretation of the treatment effects in the new experiment. To prevent irreproducible claims, in such a situation, the researcher may decide to assign wider prior distributions or to include additional animal strains in the new experiment. In fact, the high standardization within animal experiments has been criticized [14, 23], in which case the consideration of additional data in the definition of prior distributions or the interpretation of the results could be beneficial. In any case, the predictions for the new experiment should be made under consideration of the assumed relatedness of the historical and new experiment's data. With internal historical data, a great similarity between the historical and the new data may be expected. With external historical data, in contrast, the animals are expected to be less similar [32, Section 3.16]. Different prediction models are further discussed below.

Although the MPD data were not directly included in the Bayesian model for the historical data, the data were still considered for setting up the prior distributions in the model of the internal data. More specifically, for the model of the outcome relative bone volume, the MPD mice were used for defining a priori plausible prior distributions for the overall mean and the residual SD. Choosing the prior distribution independently of the data respects the Bayesian ideal of the prior distribution as additional information to the data. With this approach, on the one hand, a larger amount and variety of information is included in the analysis, resulting in a more general prior distribution [147]. On the other hand, a prior from wild type animals may bias the posterior and thus the predictions for the new experiment. As an alternative to defining prior distributions independently of the modeled data, the prior distribution of the historical model may also be centered around estimates of the internal data itself in an empirical Bayes approach [99, 100]. This reduces the amount of information that is contained in the approach but also reduces the risk of bias. The results from an empirical Bayes approach are generally expected to be more similar to a frequentist approach (since no additional prior information is included), but the empirical Bayes approach still maintains most of the advantages of Bayesian analysis, such as a better representation of uncertainty and the ability to derive a predictive distribution as prior for the new experiment.

For the outcome bending stiffness, it was challenging to find measurements in the MPD that are comparable to the internal data. This could have been due to a loose notion of bending stiffness and differences in measuring machines. Yet, it remained unclear what caused the large differences in internal and external animals and as a consequence, the external data were not

used for prior specification of the historical data model for that outcome. To get a greater benefit from available historical data, it is therefore important to not only collect historical data but also to set up common definitions of the measurements (ontology, such as the Mammalian Phenotype Ontology [137]) and to standardize measuring procedures. The topic of animal databases is further addressed in the outlook (Section 4.4).

To demonstrate the potential of the here proposed methodology in a scenario where data from multiple (possibly external) experiments are available, a small fictive historical data set of five historical experiments, consisting of two groups (control and experimental) was simulated. For both the simulated and the real-world data, it was assumed that the data follow a normal distribution and interest lies in the difference in the control group mean and experimental group mean (the case of non-normal data is discussed below). For the simulated multi-experiment data, heterogeneity between the experiment-specific means was modeled by normally distributed deviations from the common mean. The heterogeneity in the historical experimental treatment groups was modeled to be greater than the heterogeneity in the control groups. This mimics the situation of a series of experiments in which the same pathological phenotype is investigated under different treatments (that are only distantly related) but with rather similar control animals. The data in the experimental group was simulated with a very large effect according to the classification by Cohen [148] and its extension by Sawilowsky [161]. In general, very large effect sizes are necessary to get a power of 80% with sample sizes of ten or less per group [9–11].

Bayesian models of the historical control data

The models for the new and the historical data were fitted as Bayesian regression models with the `brms` package [65] in R [66]. The package provides user-friendly and flexible functions facilitating the model fitting and handling of the results for more applied researchers with little statistical knowledge or researchers that are used to frequentist approaches. The package uses the `Stan` software [149], which implements fast Markov Chain Monte Carlo (MCMC) algorithms. MCMC algorithms allow for the approximation of complex posterior distributions and thus overcome the limitations of exact analytic computations, that for a long time hindered the practical application of Bayes methods [65]. The regression models were either fit with an intercept only or with additional "fixed" population effects or "random" subgroup or experiment-specific effects (heterogeneity). Including data from multiple animal experiments led to a Bayesian random-effects meta-analysis as a special case of a Bayesian hierarchical model. The random-effects model reflects the heterogeneity in the historical and new experiments, that may come, e.g., due to different animal housing conditions, experiments designs or animal batches [31, 162, 163].

The prior distributions for the mean parameters in the historical data model were taken from the normal distribution family, which is the conditionally conjugate family of the normal distribution family of the data (likelihood distribution). Using a prior from the conditionally conjugate family of the likelihood distribution allows interpreting the prior distribution as equivalent data [59, Section 2.5]. The SD of the priors were set to the SD of a single typical observation (unit information priors (UIPs) [60]). This requires an estimate of the SD of a single typical observation, which was based either on internal or external historical data (as outlined above). The prior for the variance parameters in the historical model were assigned half-normal distributions, as recommended by Gelman [91]. This choice allows fitting the meta-analysis model also in the situation of very few and small experiments, where frequentist methods are problematic [37]. The prior distributions for the mean and residual SD were modeled to be independent, as in most practical applications [46, 94, 95]. As an alternative, Bonapersona et al. [8] use a two-dimensional conjugate prior distribution of the normal-Gamma family. A two-dimensional distribution could also have been fitted in this thesis to the predictive distributions of the mean and residual SD for the new experiment. The disadvantages of using two-dimensional predictive distributions are, however, that they are more difficult to fit and more difficult to understand by applied researchers. Furthermore, they imply more assumptions in terms of the correlation structures of the mean and residual SD, which may not be well supported in the case of few data. Moreover, with dependence in the prior distributions for the mean and residual SD, the assumption for the calculation of the Bayes factor with the Savage-Dickey density ratio may no longer hold and more complicated methods may be necessary [164].

In this thesis, a meta-analysis model with individual animal-level data (individual animal data meta-analysis (IADMA) or hierarchical model or individual patient data meta-analysis (IPDMA) in a clinical context) was considered, as opposed to a meta-analysis model with aggregate data (aggregate data meta-analysis ADMA). The IADMA was performed in a one-step approach as a hierarchical model [165]. Alternatively, an IADMA can be performed as a two-step approach [166–168]. The reasons, why a one-step approach is chosen here, are that the one-step approach is more flexible and that, with the two-step approach, estimates of the common mean, the standard errors, and the heterogeneity may be biased in the situation of few data per group [169–171]. Moreover, the two-step approach is less suited than the one-step approach to detect non-linear associations and interactions between continuous covariates and the outcome [172–174]. If, however, no historical data are available on an individual-animal level or if the researcher decides to perform an ADMA, because they want to use the summary estimates from published literature, then the here proposed sample size planning approach is still applicable. Moreover, if the historical experiments deviate from the new experiment regarding population characteristics, that are considered to have an impact on the outcome of interest, meta-regression [175, 176] or

network meta-analysis [176, 177] may be used for fitting a model to the historical data. In each case, a Bayesian meta-analysis allows a better reflection of uncertainties in the historical data than a frequentist analysis [47]. In addition, by choosing reasonable prior distributions, Bayesian models can also be fitted in the challenging situation where the number of historical groups or experiments is very low [37] or the measurements of only one group or experiment are available (in which case the posterior distribution of the heterogeneity corresponds to its prior distribution [46]).

Predictions for the new data

Prediction models

The Bayesian regression model estimated in the historical data was used to make predictions for the mean in the control group of the new data and for the residual SD. Predictions can be made based on various assumptions on the degree of similarity between the historical and new control data [32, Section 3.16]. For the C5aR1 experiment, the predictions for the new experiment were based on internal data only. For the outcome bending stiffness, internal data in several age groups were available. Since, according to the historical data, the outcome seemed to highly depend on the animals' age, the covariate "age" was included in the prediction model for the mean. In general, it cannot be expected that population characteristics like the animal's age or sex coincide in the available historical data [178–181]. Then the modeling of these characteristics may be necessary to prevent bias in the predictions. The inclusion of the age effect induced additional variability in the prediction for the new mean. This is because the predictions were built as the sum of the MCMC draws of the overall mean and of the age effect. Unlike frequentist point predictions, this Bayesian prediction, thus, reflects a greater uncertainty with few historical data than with a greater amount of historical data, and this uncertainty is incorporated in the sample size calculation. There, a greater uncertainty leads to wider prior distribution in the new experiment and generally requires a larger sample size to obtain a prospective target power. For the simulated multi-experiment data, predictions for the mean were made from a meta-analysis model. This allowed the new experiment's mean to deviate from the historical means because of heterogeneity. As a further extension, instead of assuming similarity of the historical and the new SDs, the prediction for the SD may as well be derived from a random-effects model [46] or under consideration of covariate effects.

A Bayesian random-effects meta-analysis model could have been fitted also in this case of the C5aR1 experiment with only one single historical experiment. In this case, the posterior distribution of the heterogeneity parameter in the model of the historical data model would have been

the same as its prior distribution [46]. The random-effects model would result in a wider variance in the posterior predictive distribution of the mean and residual SD, as compared to the model under the assumption of identical parameters in the new and historical experiment. This would in general lead to a larger sample size and be a more conservative method. However, this case requires special care in the prior definitions. However, the adequacy of the heterogeneity prior distribution can hardly be judged, when there is no information in the historical data on these parameters. Prior distributions that are too wide may result in predictive distributions for the new experiment's control animals that are too vague, and thus limit the value of a Bayesian approach to the design planning and analysis.

Approximation of the predictions by parametric distributions

MCMC approximations are a general and flexible method for approximating Bayesian parameter distributions. However, when it comes to the definition of prior distributions for the new experiment, the representation of the predictions by series of large numbers of MCMC draws is inconvenient. To transform the MCMC representations into a format, that can be interpreted and handled more conveniently in the planning of and analysis of a new animal experiment, the distributions of MCMC draws were approximated by parametric distributions.

A normal distribution was fitted to the MCMC draws of the predicted mean parameter and a log-normal distribution was fitted to the draws of the historical SD using maximum likelihood (ML) estimation. Alternatively, an inverse-Gamma distribution could have been fitted to the square of the residual SD (as residual variance) [8, 46], which represents a conditionally conjugate distribution of the normal likelihood distribution of the data [59, Section 2.6]. A log-normal distribution for the residual SD was chosen as prior instead because this distribution is assumed to be easier to interpret for the applied researchers conducting the animal experiment, as it allows modeling the logarithm of the residual SD as a commonly familiar normal distribution. Furthermore, the prior distribution was derived for the residual SD instead of its square, the variance, because this allows a more direct interpretation [91]. In normally distributed data, the empirical variance follows a Gamma distribution. A log-transformed gamma distribution, in turn, can be approximated by a normal distribution. This approximation can be used to realize the parametric approximation of the logarithm of a predicted variance from an ADMA (as described above) by a normal distribution [182], similar to the approximation that was here performed with individual animal data. A problem of this approximation of the empirical variance from aggregate data is, however, that the approximation accuracy of the empirical variances in the aggregated data depends on the sample sizes of the historical experiments. Hence, with little data in the historical experiments, more accurate results may be achieved by using individual-level animal data, like in this thesis.

Instead of fitting one component parametric distributions, mixture distributions consisting of multiple components (as a representation of a collection of cumulative distribution functions) could be fitted to the MCMC draws [84]. An advantage of mixture distributions is that they are more flexible and can thus fit the predictive distributions more closely. For example, by the representation of the prediction of the mean or the logarithmic SD by multiple normal distribution components, mixture distributions allow focusing an essential part of the total probability mass on the mode of the MCMC draws, while modeling heavy tails by a vague normal distribution component that has a large variance [84]. This makes the model more robust to prior-data conflicts, by allowing the data to dominate the posterior distribution. The robustness can be even more increased by assigning the vague component more weight in the cumulative mixture distribution. However, in a situation where the overall sample size of the historical experiments is large, but the sample size in the single new experiment is low, it may not be desirable for the posterior distribution to be very sensitive to the likelihood distribution of the new data, as further discussed below. Moreover, a parametric fit of a simple normal distribution is easier to interpret and discuss with the responsible animal researchers and is more in line with how a researcher would define a priori expectations on the distribution of a parameter. Furthermore, simpler distributions are less likely to overfit the distribution. The fitted one-component distributions seemed to capture the central tendency and spread of the distributions adequately, although not as closely as mixture distributions. If a more flexible model is nonetheless desired, the suggested sample size procedure may still be conducted, but with mixture distributions as parametric approximations of the predictions instead of single-component distributions.

The parametric distribution, that was fitted to the residual SD of the historical control animals, was used as prior for the new experiment's residual SD in both the control group and the experimental group. Alternatively, a predictive prior could have been obtained from historical data with a treatment that is expected to have similar effects to the treatment in the new experiment. However, no actual knowledge on the specific treatment for the new experiment is expected to be available for a new treatment and with only little data from a pilot experiment, the estimates from single small experiments are likely to be very unreliable. It was argued that there is likely to be more information available on the control groups since the control groups from multiple experiments on the same phenotypic outcome can often be expected to be similar [8]. If it is expected, that the new experimental group's residual SD substantially deviates from the control group's SD, an alternative prior distribution may be formulated, as here demonstrated in a sensitivity analysis. Moreover, if the experiment's researcher wants to reflect greater uncertainty in the prior of the residual SD they may assign the prior a larger variance parameter or fit a mixture distribution in which the vague component is assigned a larger weight (as explained above).

Comparison of meta-analytic predictive and meta-analytic combined approaches

In this thesis a meta-analytic-predictive (MAP) approach was applied, using a meta-analysis model of the historical data to formulate a predictive prior distribution for the Bayesian analysis of the new experiment [39]. An alternative to this sequential approach for the combination of historical and new data is a meta-analytic-combined (MAC) approach. In a MAC approach, both historical and new data are analyzed in one step in a common Bayesian model. Following a MAC instead of a MAP approach for the planning and analysis of a new animal experiment would come with the advantage of being more direct and easier to implement. This is because the MAC approach requires only to fit one single Bayesian model, instead of both a model for the historical data (which is used for derivation of a MAP prior from the historic data) and a model for the analysis of the new data. In theory, under the assumption that the parameters in the historical and the new experiments are exchangeable, the MAP approach is equivalent to the MAC approach [84]. This means, that the two approaches should yield on average the same posterior distributions in the analysis of the new experiment's data. However, this theoretical equivalence refers to the analytic situation where the exact predictive distribution, that is obtained from the historical model, is used as prior distribution, instead of an approximation. In the case of model fitting with MCMC methods, the posterior distributions are approximated and no analytic solution prediction is available. With a close fit of a parametric distribution to the predicted distribution, the equivalence of the MAP and MAC approach can be maintained and the sample size for the here applied MAP approach should be the same as the sample size from a MAC approach, where the historical and new data are analyzed in a common model. However, when the posterior predictive distribution is approximated less closely, the MAP approach may result in wider or narrower posterior distributions and in turn in smaller or larger sample size results than in a common analysis of all data in a MAC approach.

To perform sample size calculation in a MAC approach, data could be simulated under varying designs, similarly to data fitted here. In the MAP approach, the prior distributions and parameters of the data likelihood distribution for the design simulation were based on a Bayesian meta-analysis model of available historical data. Instead, in a design simulation with a MAC approach, the prior distributions and model parameters could be based on a frequentist meta-analysis of the historical data. This would correspond to the prior definition in the Bayesian model of the historical data in the preceding step of the MAP approach. An example of sample size planning with a MAC approach is given by Qi et al. [183] (as an extension of the sample size planning procedure by Sutton et al. [184]). In the approach by Qi et al. [183], hypothetical new data are simulated from the MAP prior for the mean, which is derived from a meta-analysis model of the historical data. Then, Bayesian models are fitted to combinations of this simulated data

and the historical data in a MAC approach. In the context of design analysis and sample size determination, however, the MAP approach has the advantage of allowing in advance judgment and control of the influence of the historical data on the estimation results in the new analysis. For this reason a MAP approach was used in this thesis, although a similar sample size planning procedure concept could be set up with the MAC approach as well.

4.1.2 Sample size determination for the new experiment via design simulation

Prior distributions for the Bayesian analysis of the new experiment

The sample size calculations were performed in a design simulation context by repeatedly simulating data of a control and an experimental group with varying differences in the means δ and estimating the power for hypothesis tests in a Bayesian analysis. The Bayesian analysis was performed as completely Bayesian analysis, where prior distributions are assigned to all model parameters (which are here the treatment group means and residual SDs). As prior for the parameters in the control group, a normal and a log-normal distribution were used for the mean and SD with ML estimated parameters, that resulted from approximating the predictions obtained from the historical data, as described above. As prior for the mean parameter in the experimental group, a normal distribution with the same mean parameter as in the control group and a SD, that corresponds to the residual SD of a typical observation, was used as recommended default prior distribution [60]. The prior distribution for the mean in the control and the experimental group were chosen to have the same expected value, since otherwise the prior distribution in the experimental group may bias the Bayesian test decision of whether there is a difference in the treatment means. The larger variance parameter of the prior distribution for the mean in the experimental than the control group in turn reflects the fact, that the prior in the control group is based on evidence from historical control data, whereas for the experimental group, no data on the treatment of interest was assumed to be available. The prior for the residual SD in the experimental group was chosen to coincide with the prior for the residual SD in the control group. Alternatively, a less informative prior distribution in the experimental than in the control group could have been chosen for the SD parameter as well. However, since the comparison of main interest was the comparison of the posterior mean parameters, the prior distribution for the SD parameters was only secondary interest.

As an alternative to a completely Bayesian analysis, the (simulated) data of the new experiment could be analyzed in a partly Bayesian analysis, where the residual SD is modeled as a fixed value, as in the Bayesian MAP approach by Neuenschwander et al. [39]. This fixed value can, for example, be a point estimate from the historical data. In such a partly Bayesian analysis

the posterior distributions are easier to calculate since exact analytic formulas are available [59, Section 2.5]. In contrast, the posterior distributions in the completely Bayesian analysis have to be calculated as marginal distributions (as averages over the parameter space of the additional model parameters). On the other hand, point-estimates of the residual SDs, that are derived from historical data, are often imprecise [7] and the uncertainty is only reflected in a completely Bayesian analysis. Compared to the impact of the prior distribution for the mean parameters, it is known that the distribution of the SD has only little influence on a Bayes factor that refers to the difference in the group means [115]. Since the Bayes factor was used for hypothesis testing, it is expected, that also the impact of the SD on the hypothesis test result will be small. On the other hand, it has been recommended to use informative prior distributions for all model parameters, when testing hypotheses with the Bayes factor [87]. Moreover, it has been noted that the role of the prior distributions becomes especially relevant in the scenario of small group sizes [115], which is the case in translational animal research [7]. There, prior distributions can reflect a priori uncertainty in the model parameters. In contrast, the use of fixed estimates is delusive, because it does not account for the actual amount of information on the parameters. To investigate the effect of modeling the residual SD as a random variable on the sample size decision, the power and type I error rate, that resulted from the completely Bayesian design in the planned experiment, were compared to the values, that resulted from a Bayesian model with fixed SD. The fixed SD was set equal to the median of the predicted residual SD from the historical data model.

Definition of the effect size and means in the new experiment

The data for the new experiment were simulated as random draws from two independent normal distributions. Each normal distribution has two parameters, a mean and a SD. For the control group, the mean parameter was chosen to coincide with the expected value of its prior (that was estimated from the historical data). For the experimental group, the mean was set to a sum of the control group's mean plus a fixed effect (as mean difference) δ. The mean difference was specified based on a hypothesized minimal clinically relevant effect δ^*. Then, for each candidate control group sample size, the power of the Bayesian tests was estimated by simulating data with the specified minimally relevant effect size δ^* and the type I error rate was estimated by simulating data with equal group means. In practice, δ^* may be specified by the new experiment's researcher as an expert in the research field. The problem in the specification of the effect size is, that in the planning stage of the design, the size of the true mean difference between the future experimental and control group is unknown. However, the specification of an effect size is necessary for power calculations and experimental planning. Still, even in the absence of evidence for the experimental mean difference in vivo, a minimal biologically relevant effect

may be specified based on which change in condition, as compared to the pathological control group, the researcher considers to be minimally necessary for the new treatment to appear useful enough for further investigation. The decision should be made in light of available knowledge on the variability in the experimental outcome [12]. In some situations, it may also be necessary to specify mean differences based on what sample size is feasible regarding constraints on the research resources [185].

A critique on the approach of determining the mean difference based on some minimal clinically relevant effect is, that the determined value may be over-optimistic [12], e.g., due to the researchers' opinion on the effect being biased by overestimated mean differences from the literature [186–188]. This in turn may lead to underpowered experiments [8]. However, the overestimation risk decreases if the minimally relevant effect is evaluated in light of available evidence on the variability in the experimental outcome. A common alternative approach for specifying an effect is to use estimates from the observed differences in single historical experiments. This empirical approach, however, is likely to lead to the specification of unrealistic effects [7]. First, the effect estimates from historical experiments are likely to deviate from the treatment effect in the new experiment, since the treatments in historical experiments typically deviate from the treatment, that shall be investigated as candidates in a new experiment (although the control groups may be similar). Data, that suits the treatment definition of the new experiment, may be available from pilot experiments. However, pilot experiments are often small, and their estimates are thus imprecise. Furthermore, taking effects from small experiments, where the treatment effects have turned out to be significant, is likely to lead to overestimation of the magnitude of the effect, since in those experiments the exaggeration of the effects is usually high [12, 189]. This in turn leads to sample size estimates for the new experiment that are too low to reach the predefined statistical goals and to a decreased probability that those effects, for which statistical significance was reached in the new experiment, correspond to the true effect.

Definition of the residual SDs in the new experiment

In the previous paragraphs, the choices for the mean parameters in the simulation of the new experiment were explained. Depending on whether the power or the type I error rate was estimated, the data were simulated with unequal or equal means. In each case, the mean parameter in the simulated control group was chosen to coincide with the mean parameter of its prior (which was predicted from the historical data). The SD parameter in the simulated control group was chosen similarly. However, instead of setting the SD to the mean of its predicted prior distribution, the SD of the control data was set to the median of its predicted prior distribution. This is because the predictive distribution of the SD was modeled by a log-normal distribution, which is skewed,

and the median is less sensitive than the mean to extreme values. In the basic design simulations, the SD in the experimental group was set up to coincide with the SD in the control group (homoscedasticity), under the assumption that no other information was available. If the experiment's researcher wants to examine which power would be necessary under heteroscedasticity, it was suggested to simulate data under varying degrees of heteroscedasticity and check if the sample size, that was chosen under a homoscedasticity assumption, would be large enough also for these designs. If the researcher finds a certain degree of heteroscedasticity realistic, but the chosen sample size does not seem sufficient for this case, an increase in sample size is suggested.

In all results, the sample sizes resulting from the Bayesian design analysis approach were constrained to be at least three, since with fewer animals, statics cannot be meaningfully estimated. More precisely, a sample size of two would be the lowest number for which SDs could be calculated, and this number may be used as an alternative minimum. The sample size for the experimental group size was kept at the value that was calculated for the frequentist test (with the estimated SD from the historical data) since it was assumed that no adequate historical data for this group were available. If such data exist, prior distributions may also be derived from that data, as stated above, and a reduction in sample size may also be considered for the experimental group.

Choice of the Bayesian hypothesis test

Considered Bayesian hypothesis tests in this thesis

For a specified mean difference and with SDs set to the medians of the predicted SD from the meta-analysis model, the necessary sample size was first calculated with the frequentist two-sample t-test. Then, design simulations were performed with this sample size as a starting point. The simulated data were evaluated with Bayesian methods that include as prior information the predictions that were obtained from the meta-analysis model of the historical data. For each simulated data set, a Bayesian hypothesis test was performed. The Bayesian hypothesis tests were either based on the Bayes factor and the posterior equal-tailed interval (ETI).

Hypothesis tests with the ETI were defined as whether the ETI included the value null. As an alternative to hypothesis tests based on ETIs, hypothesis tests could be performed with highest posterior density intervals (HDIs) as a further type of Bayesian credible interval, that always includes the mode(s) of the posterior distribution instead of its median [190]. The differences in the definition of ETIs and HDIs relate to their practical use [1]. When an ETI is constructed for

skewed distributions (such as the log-normal distribution that is used for modeling the predicted SD for the new experiment), values of the parameter distribution that lie inside the interval may be less probable than parameter values that lie outside the interval. This is not the case for a HDI. Moreover, the HDI has the smallest width of all possible credible intervals, which is a desirable property when it is used for posterior inference. On the other hand, advantages of an ETI are, that it is very simple to obtain an ETI from the sample quantiles of the posterior distribution. In addition, when parameters are transformed before the calculation, the intervals of the parameters on the original sale can be obtained by simple back-transformation of the ETI bounds. In symmetric distributions, as obtained for the mean parameters, the quantile interval and the HDI return similar results. In this thesis, ETIs are used to yield comparable results with the Bayesian approach for animal experiments that was proposed by Bonapersona et al. [8]. Results with HDIs were provided in my previous contribution [1].

Hypothesis tests with the Bayes factor were defined as whether a threshold for significance was exceeded. The threshold was chosen to indicate at least moderate evidence for the alternative hypothesis. "At least moderate evidence" was defined in terms of the categorization scheme of Stefan et al. [56], which was adapted from Lee and Wagenmakers [119] and Jeffreys [120]. This evidence categorization scheme is convenient, but should only be supplemented with a visualization of the whole distribution of the Bayes factors [52, 120, 146], as presented in my previous contribution [1]. Additional steps in a robust workflow for Bayes factor analysis include simulation-based calibration and sensitivity analyses to ensure a correct interpretation [146]. The Bayes factor hypothesis test may further be adapted to include the decision of no evidence. A further extension may address the fact that choosing a heuristic threshold value does not guarantee an optimal choice regarding the "cost" or "values" as consequences of possible decisions (such as frequentist type I and type II errors). These consequences can be formalized in utility functions and define, for example, whether it is worse to falsely reject the null hypothesis when the true group means are equal (type I error), or to falsely not reject the null hypothesis when the true group means are unequal (type II error). In a typical clinical or translational research setting, type I errors are usually regarded as worse than type II errors, as reflected in a lower conventional value for the significance level than for the target power of a hypothesis test. In terms of a utility, this can be reflected by defining an overall utility with a higher absolute value of the utility of type I errors than of type II errors. After such a utility has been defined, the Bayes factor threshold can be optimized to get the best expected overall utility and thus a hypothesis test, that is most in line with the intended consequences of possible decisions. A problem with the definition of utility functions is, however, that there are currently no standards for defining the utility values, so further research is needed on this topic [146].

By repeatedly simulating hypothetical data with unequal group means and by calculating the Bayesian hypothesis tests, the power was calculated as averages of rejections against the null hypothesis of equal group means. Similarly, the type I error rate was calculated in simulated data with equal group means. An overall sample size of the control and experimental group was considered large enough if a "power" of at least 80% was reached and if the "type I error rate" of the corresponding design did not exceed the nominal level of 5%. Then, the number of animals in the control group was iteratively reduced over the number, that was calculated for the frequentist analysis. In each iteration step, it was checked if the power of the Bayesian tests was still large enough and the type I error rate small enough before the next step was performed.

Arguments for and against using Bayes factors for hypothesis testing

The sample size results with the Bayesian hypothesis test based on the Bayes factor (indicating at least moderate evidence for the alternative hypothesis [56]) were compared to the results from the frequentist two-sample t-test. Furthermore, the Bayesian hypothesis test results based on the Bayes factor (indicating at least moderate evidence for the alternative hypothesis) were compared to the results of a Bayesian hypothesis test based on whether the ETI of the mean difference included the value null. By definition, Bayes factors can be seen as a generalization of frequentist likelihood ratio tests for point hypothesis to arbitrary hypothesis. This enables more flexible hypothesis testing [146]. In addition, hypothesis testing based on Bayes factors (or on Bayesian approaches in general) has the advantage over frequentist testing, that it can better handle uncertainties in the estimation of parameters [146]. Furthermore, repeated evaluation of the Bayes factor may be used to resolve some limitations of classical group sequential designs, which can be useful to save animals if early stopping of the animal experiment is possible [52, 56, 191].

In comparison to credible interval-based hypothesis tests, hypothesis tests with Bayes factors have, in the situation of a point null hypothesis (like here for a mean difference of exactly zero), the interpretational advantage, that they don't violate the principle of predictive irrelevance [54], whereas credible interval based tests and frequentist tests do so. Violating the principle of predictive irrelevance means that it can happen, that data, that seem equally likely under the null and the alternative hypothesis, can influence the decision of whether the null hypothesis is rejected. The interpretation of such a situation is unclear, and such paradoxical results can be avoided by using the Bayes factor for testing hypotheses. Bayes factors quantify the evidence for both the model under the alternative hypothesis and the model under the null hypothesis. With this property, Bayes factors can also be used to distinguish a scenario, where the data do not provide enough evidence for either hypothesis, from a scenario, where the data provide evidence

for the null hypothesis [192, 193]. In the former scenario, more animal data should be collected, whereas in the latter scenario, the treatment in the experimental groups seems to be of no benefit, as compared to the control group, and further data investigation in clinical trials in humans seems futile.

As an alternative to testing hypotheses based on Bayes factors or testing if a credible interval includes a null value, Kruschke [190, Section 12.1.1] suggests testing whether the posterior interval excludes a region of practical equivalence, which may be defined as all parameter values that are smaller (in absolute value) than a minimal clinically relevant mean difference. However, the problem with using this decision rule in translational animal experiments is, that it requires large sample sizes or very wide regions of practical equivalence to discriminate between equivalence and non-equivalence of the group means [194]. As a consequence, for translational animal experiments, where sample sizes are generally small, this approach has only little power. Moreover, in a decision-theoretic framework where utility functions (as defined above) are formalized in terms of expected posterior loss, a critique on the region of practical equivalence decision rule of Kruschke [190, Section 12.1.1] is that, when applied with the HDI, this decision rule is not in line with actual real-world purpose of interest and, also when applied with the ETI, the implied utility value is only arbitrary and thus hard to justify [195]. Furthermore, Kelter [53] shows in a simulation study that the Bayes factor is more sensitive to detect an existing effect than the 95% region of practical equivalence decision rule and that it should be preferred over the classical frequentist p-value based decision rule. For these reasons, the region of practical equivalence decision rule was not further considered here.

Alternative approaches to sample size determination

Frequentist vs. Bayesian approaches to the design simulation

While the model was fitted in a Bayesian framework, the hypothetical data in the new experiment were generated in a frequentist framework with fixed values for the group means and SDs. Hence, randomness in the simulated data came only from the normally distributed residuals. However, the analysis differed from a frequentist analysis, since uncertainty in the parameters was incorporated through their prior distributions. An alternative approach to this data simulation would be a simulation, where the data is simulated with random means and SDs. For the control group, this could be realized by sampling in each iteration a value for the control mean and SD from the (MCMC or parametric) predictive distributions. Then, the data is simulated with these parameters. After the hypothesis test is performed, a new mean and SD would be sampled from the predictive distributions and the second hypothesis test would be performed in a new data set,

that is sampled with new mean and SD parameters. Thus, the parameters of the likelihood distribution would vary, while the prior distributions would stay the same. For the experimental group, the parameters could be generated by first sampling a value from the control mean's predictive distribution and then adding a random effect δ. Under the assumption of homoscedasticity, the SD could be sampled from the control SD's predictive distribution. In an additional analysis (that is not presented in this thesis for parsimonious reasons) such a Bayesian data simulation approach with random means and residual SDs of the simulated data was investigated. However, this data generation led to unrealistically large uncertainties in the posterior distributions and more meaningful results could be obtained by incorporating the uncertainty in the analysis only by the prior distributions in the Bayesian analysis. Furthermore, since the hypotheses were formulated with fixed parameters (as they should to guarantee a correct interpretation of the Bayes factors [104]) it may seem logical to use fixed values for the data generation as well and represent the parameter uncertainties only via the prior distributions in the Bayesian analysis.

Setting sample sizes via design simulation vs. via an effective sample size

As an alternative to either frequentist or Bayesian design simulations, the sample size of a new experiment may be determined by reducing the number of animals, that was calculated from a frequentist power calculation, by the estimation of an effective sample size (ESS) of the predictive prior for the mean [39, 84, 196, 197]. In one-parameter models (or conditionally on fixed values of the additional model parameters) and using conjugate prior distributions (as defined above), different methods for estimating the ESS yield the same result [107]. However, with non-conjugate prior distributions, the ESS estimates may vary depending on the estimation approach [107] and with the residual SD as an additional parameter (or further covariates), the calculation of an ESS becomes more complicated [183]. A prior ESS for two-parameter models has been proposed [198]. However, this concept only applies to a specific situation. In this thesis, design simulations have been used for sample size calculation, since they can be applied in very general situations. They allow designing experiments to reach a specific statistical power in the future Bayesian analysis of the new data and also the investigation of additional design quality criteria (as discussed below). In contrast, the sample size, that is calculated via the ESS, is based only on frequentist analysis, without consideration of Bayesian power or other design quality criteria, and does not consider the possibility of prior-data conflicts [183].

Examination of additional design quality criteria

As discussed by several authors [121, 124, 146, 190], there exist more meaningful goals than power in the Bayesian context. For example, a more meaningful goal might be to aim for sample sizes, that promise a certain precision in terms of achieving a high probability of the posterior credible interval of the effect (here the mean difference) being smaller than some desired threshold [84, 121, 124, 190]. Furthermore, Gelman et al. [36, 124] argue that judging a model's quality based on other model characteristics like the type S (sign) error rate, the exaggeration factor and the mean square error (MSE) are more meaningful than aiming only for the highest power. The type S error rate is the probability that an estimate of the true effect has the wrong sign, given that is statistically significant. The exaggeration factor is the ratio by how much an estimate is greater in absolute value than the absolute value of the true effect, given that it is statistically significant [12]. Especially in small data sets, like in the groups of translational animal experiments, and with smaller true effects the estimates are characterized by a high uncertainty and the exaggeration factor and the type S error rate can be large [124, Section 4.4] [1]. This leads to a high probability that statistically significant estimates are exaggerated or even in the wrong direction. The problem of exaggeration is aggravated by publication bias, where only the statistically significant findings are published but not the non-significant findings, since then only the highly exaggerated effects become published. This, in turn, may cause many overestimated effects in published literature and may contribute to the failure to reproduce published findings (reproducibility crisis) in animal experiments [15, 18, 20–23, 199–201]. For these reasons, additional quality criteria such as average effect estimate, coverage probability and average width of the posterior credible intervals of the effect (as mean difference in the treatment groups), exaggeration factor, type S error rate, and MSE were examined for the calculated sample sizes and their corresponding mean differences and were compared between frequentist and Bayesian analysis with informative prior distributions or with UIP.

Factorial designs, multiple endpoints and normality assumption

In the basic simulation model, equal group means of the new experiment's control and experimental group were assumed. To investigate the effect of heteroscedasticity on the power with a chosen sample size, data were simulated with that sample size and various degrees of heteroscedasticity. Then, a Bayesian distributional model was fit. In a distributional model, both the mean and residual variance are allowed to vary by group [78], generalizing the model to suit a broader spectrum of experimental designs. As a further generalization, the case of non-normal data can be considered. All the hypothesis tests and models in this thesis were based

on the assumption, that the measurements of the experimental outcomes follow a normal distribution (possibly after some transformation). With small sample sizes, the normal assumption is often hard to verify and non-parametric methods have been suggested as an alternative, such as the Wilcoxon(-Mann-Whitney) test [202] for the comparison of two independent samples in a frequentist framework. A Bayesian counterpart to the Wilcoxon test has been proposed by van Doorn et al. [203]. Alternatively, non-parametric approaches can be used in a Bayesian framework to weaken the exchangeability assumption, that was made for the historical and the new experiments [204, 205]. The here proposed approach for estimating sample sizes in Bayesian predictive frameworks based on historical controls can also be extended to address non-normal data like ordinal and censored data using the `brms` package [65, 206]. Furthermore, the analysis and test results can be made more robust to extreme observations and to data deviations from the model assumptions by using prior distributions with longer tails (for example, normal prior distributions may be replaced with t-distributions [59, Chapter 17]) or by using mixture distribution with vague components, as explained above.

In the practice of translational animal experiments, the designs and hypothesis tests of real-world animal experiments may be complicated by repeated measurements, multiple subgroups or time points or due to the measurement of multiple experimental outcomes [6]. When testing confirmatory hypotheses in such experiments, problems of multiple comparisons occur and the control of the type I error rate may become challenging [6, 207]. To adjust for multiple comparisons in a frequentist context, the significance level is adapted, which generally leads to a larger required sample size. Typically, animal experiments rather have an exploratory than a confirmatory character [6, 16], but the distinction is not always clear [7, 16]. Piper et al. [6] argue that, in exploratory experiments, hypotheses can be tested without a correction for multiple testing, if the results are interpreted with the exploration character in mind and if the results are published with a likewise indication. In addition, Gelman et al. [208] argue that, in a Bayesian framework, multiple comparison adjustments are usually not necessary, since the risk of making wrong conclusions is addressed in Bayesian hierarchical models, where estimates are automatically shifted towards each other.

To test if there are differences between treatment groups in any of multiple experimental outcomes, multivariate analyses could be used. Like testing, for a specific experimental outcome, whether there are differences in any pair of groups (as a generalization of two-group comparisons), testing whether there are differences between two groups in any outcome can be addressed in "global" or "omnibus" hypothesis tests. However, such global tests often don't address the questions of the experiment's researchers directly [209, p.238]. This makes it often necessary to conduct further testing of pairwise differences in pairs of groups for single outcomes in "post-

hoc" analysis. To address the researchers' questions more directly, it may, thus, be preferable to perform the two-group comparisons for single outcomes directly. Therefore, the here illustrated case of two-group comparisons in univariate analysis was considered as most relevant, although the experiments may involve multiple outcomes and multiple group comparisons. In principle, however, the methodology may as well be extended to more general designs and the test thresholds may be adjusted for more strict error control.

Investigation of prior-data conflicts

In the context of designing and analyzing experiments with Bayesian predictive approaches prior-data conflicts [38, 61–64] may occur if the estimates, that are derived from the new experiment's data, appear not very probable according to the predicted prior distribution from the historical data model. Such a conflict may, for example, result from an incorrect model of the relationship between the new and the historical data, or from bias in the historical or new data. Then, the conclusions based on Bayesian hypothesis tests or parameter estimates may (strongly) deviate from frequentist conclusions that are based on the observed data from the new experiment only.

Methods exist for testing whether there seems to be prior-data conflict, once the data in the new experiment are observed (e.g. [38, 61, 62, 210]). These checks use the newly observed data to judge the plausibility of the prior distributions. Ideally, the data should be used to check first if there is any disagreement with the statistical model, that was assumed for the data, and then proceed to check disagreement with the prior distribution [61]. The intuition behind this procedure is, that checking the prior distribution is only meaningful if the statistical model for the data is appropriate. Based on prior-data conflict checks, the researcher may then decide to use a different prior for the analysis of the new experiment than the informative prior, that was chosen in the planning phase of the experiment. Possible alternative prior distributions are, for example, the choice of again a predictive prior from historical data, but with re-weighted experiments based on their deviations from the new experiment [64, 211] or less informative or more robust prior distributions with larger tails [62, 64, 84, 212]. As discussed above, the prior distributions could be made more robust by the use of t-distributions or the use of mixture distributions with a larger weight assigned to the flat mixture components. With the use of more robust prior distributions, prior information is discounted or discarded more rapidly in the case of prior-data conflicts, leading to a larger impact of the data [84]. Alternatively, the researcher may decide to analyze the data with frequentist methods instead of Bayesian methods. However, the problem with such approaches is, that statements about the true values in the new experiments' data can only be made based on their empirical estimates. This is problematic in the context of translational animal experiments, where the small sample sizes are small, and especially in the

situation where the number of control animals is reduced by Bayesian analysis with informative prior distributions because with low sample sizes, the empirical estimates and checks for prior-data conflict are likely to be unreliable [8].

To decrease the risk of prior-data conflicts already in the planning stage of the new experiments (instead of only in the later analysis stage), Bonapersona et al. [8] suggest accessing the sensitivity of the predictive prior distribution to specific subgroups or experiments (like a comparison between a prior derived from both external and internal data with a prior that was derived from internal data only). Furthermore, they suggest excluding subgroups or experiments that seem suspicious or to exclude or down-weight them if their impact on the prior is considered too large. Here, prior-data conflicts have been investigated in terms of the true means or residual SDs in the simulated new experiment deviating from the expected values of the corresponding prior distributions. First, such deviations were investigated as the true mean parameters deviating from the expected value of their prior distributions. Second, deviations between the true SD parameter and the expected value of its prior distribution were investigated. Third, the simultaneous deviation of the true means and residual SDs from the expected values of the prior distributions were investigated. The results are discussed in Section 4.2.3.

4.2 Discussion of the results

4.2.1 Sample size results

In the analyzed C5aR1 example the control group sample size, which would be required with the frequentist two-sample t-test, could be reduced by the inclusion of historical information in a Bayesian analysis. This means that the goals of a target prospective power of at least 80% and a type I error rate below the frequentist nominal level of 5% could be achieved also with a smaller sample size than the minimal sample size in the frequentist case. More specifically, for the outcome bending stiffness, the size of the control group could be reduced with the Bayesian analysis to four animals, while the frequentist control group sample size was ten. For the outcome, relative bone volume, where the sample size calculation with the frequentist t-test resulted again in ten animals, the control group sample size could be reduced to three.

The overall sample size for the new C5aR1 experiment then depends on whether the researcher wants to be able to detect the specified minimally relevant effects for both experimental outcomes and whether both the experimental outcomes can be investigated in the same animals. If being able to detect the specified minimally relevant effects for the relative bone volume is considered a sufficient indication for an improvement in the bone quality, then the smaller sample size of three

animals in the control group is considered sufficient. However, for the outcome relative bone volume, a possible age effect was not considered in the prediction for the new control mean, which may also have led to a more informative prior distribution and to a lower sample size than for the bending stiffness outcome, where the age effect was considered in the prediction model. As a more conservative option or if the researcher wants to be able to detect the specified minimally relevant effects for both experimental outcomes, the larger sample size of four animals in the control group should be chosen.

In the example with the fictive historical data from multiple experiments, the results from a Bayesian meta-analytic predictive approach were compared to the sample size estimates, that resulted from the common approach of frequentist sample size calculation based on single historical experiments (instead of sample size calculation from predictive meta-analysis estimates in a joint analysis of the historical experiments). There, the frequentist sample sizes were calculated based on the means and SD of each historical experiment separately. The means in the control groups were modeled to be quite similar and the means in the experimental groups were modeled to be less similar, as described above in Section 4.1.1. Although the residual SD was modeled to be the same in the historical control and experimental groups, the empirical estimates in the small data sets differed. Since in practice, the true values are unknown and so the sample size calculation is based on empirical estimation, the sample size calculation for the separate historical experiments was performed under the assumption of heteroscedasticity with the Welch test instead of the t-test. Besides the choice of the Welch test as statistical test function, the same target power, type I error rate, and the same allocation ratio as before were used. The sample size estimation in the separate experiments led to up to 344 animals per group. In contrast, with the true values, that were used for simulating the data, a sample size of nine animals would be sufficient. This result illustrates that the sample sizes, that are calculated in a classical frequentist framework based on mean and SD estimates from a single small historical experiment, are very unreliable. This limitation is in line with the limitations of formal sample size calculations that were addressed by Mayer and Muche [7]. Moreover, the sample sizes, that were calculated in the single historical experiments varied widely, due to the variation of the estimates in the historical data.

To investigate if a better planning foundation can be obtained by using the meta-analytic predictions from a joint analysis instead of estimates from the single historical experiments, the calculated sample sizes were compared to the sample size that resulted from such a joint analysis of the historical data. For the new experimental group, instead of using estimates from the rather heterogeneous historical groups, the mean was defined by a clinically relevant effect (as discussed in Section 4.1.2), and the SD was assumed to coincide with the control group. It

was argued that this would in general represent a better approach than using the estimates of historical treatment groups, that differed in the investigated experimental treatment and are thus expected to be not very representative for the new treatment. Using the meta-analytic predictions from a joint analysis led to a sample size estimate of eight animals per group with the frequentist t-test. This is much closer to the sample size of nine animals, which results with the calculation based on the true parameters of the historical data. The better planning foundation with the joint analysis of the historical control data did not only benefit from the greater amount of historical data but also from using a minimal biologically relevant effect size and ignoring the data from the experimental animals. Still, the estimates from the meta-analysis were much closer to the true parameter values than the estimates from the single historical experiments were on average. This indicates that the benefit does not solely result from ignoring the historical experimental data, but from the better estimation of the true parameters in the historical control data in a joint analysis.

Furthermore, it was investigated if this sample size, which resulted from using the predictions from a joint analysis of the historical data as estimates in a frequentist calculation, could be reduced with a Bayesian analysis of the new data. This implies, that the predictions are not only used as estimates for the planning of the new design but also in the analysis of the new design, where they are incorporated as prior distributions. With this Bayesian predictive approach, the control group size could be reduced to three animals over the sample size of eight animals, which was calculated with the frequentist t-test. Moreover, with the Bayesian analysis of the historical data uncertainties in the mean and residual SD of historical data could be quantified and the information on those parameters could be incorporated into a Bayesian analysis of the new experiment.

While a sample size of three or four animals may appear too low from a frequentist perspective, the Bayesian analysis can compensate for the low sample size by the additional historical information. This implies that the animal experiment has to not only to be planned but also to be analyzed with Bayesian methods. If instead the experiment was planned with the lower number, that was obtained from the Bayesian approach, and analyzed with a frequentist approach, then the sample size would be too low to reach the target power. Instead of reducing the sample size in the future experiment, the additional information that comes with a Bayesian analysis could also be used to increase the power (and thus, by definition, the probability of detecting truly present effects) of the Bayesian test in the new experiment over the power of the frequentist test. As a further alternative to a sample size reduction, the advantage of a higher precision in the Bayesian predictive approach could be used to obtain a target power of 80% also in designs where lower effect sizes are of interest, as discussed below.

Sensitivity of the results to the choice of the hypothesis test

In the analyzed examples, the sample size results with the Bayes factor decision rule were the same as the sample size results with the ETI based decision rule (as defined above). However, in general, this is not true. Lovric [213], Wagenmakers et al. [54] and Kelter [53] demonstrate cases where there are contradictions between the conclusion drawn from credible intervals and from Bayes factors. More specifically, there are situations where the posterior credible interval excludes the value hypothesized by the null hypothesis, but the Bayes factor supports the null hypothesis (type I conflict), while there are also situations in which the value hypothesized by the null hypothesis lies in the credible interval, but the Bayes factor rather supports the alternative hypothesis (type II conflict) [213]. It has been recommended to report both the Bayes factor and credible intervals [214], but this may lead to confusions in such conflicting situations and may raise questions on which conclusion should be drawn regarding whether a null hypothesis should be rejected. As an approach to reconcile Bayesian hypothesis testing and parameter estimation, Kelter [215] suggests to use Bayesian evidence interval and evidence values. This may be a further extension of the here proposed methodology if experimental researchers wish to prevent the occurrence of conflicts in the conclusions from Bayes factors and credible intervals.

Sensitivity of the results to the definition of the prior distributions

Sensitivity to the prior distribution for the control mean

Simulating hypothetical data for the new experiment in a design evaluation framework allowed for investigating the sensitivity of the candidate design's assumptions to the mean difference, to the residual SD, and to the assumptions on the prior distributions. The power with the informative predictive prior distributions for the mean and residual SD was compared to power with a less informative prior for the mean in the control group as UIP, defined above in Section 4.1.1. For the mean in the experimental group, a UIP was used, as before. Using a UIP for the control mean instead of the more informative mean, which was predicted from the historical data, led to a loss in power in the Bayesian designs. More specifically, the power dropped to a value, that was similar to the power obtained for the frequentist analysis with the t-test and without any prior distributions. This indicates that, if the aim is a reduction of sample size as compared to the frequentist analysis, a more informative prior distribution than a UIP is necessary in the Bayesian analysis for the parameter, that is addressed in the hypothesis of interest (here the group mean difference).

The prior distribution of the mean difference is obtained as the difference of two random variables, which are the groups' means in the experimental and the control group. Hence, by increasing the informativeness of one of the group means, also the informativeness of the mean difference will be increased. The informativeness of a prior distribution refers to the distribution's width as compared to the width of the likelihood distribution of the data, which determines its relative impact on the posterior distribution. Typically, the informativeness of a prior distribution increases with a decreasing variance of the distribution. It was assumed that no historical information was available for the experimental group, so the informativeness can mainly be increased through the control mean. The prior distribution for the new experiment's control mean was obtained as a prediction from the posterior distributions of the historical data model. As discussed above, the variance of a Bayesian posterior distribution is determined by both the variance in the estimates from the observed historical data (standard errors) and the variance in the prior distributions of the model parameters. Increasing the sample size in the historical data will generally decrease the standard errors in the estimates from the data and increase the data precision. Without consideration of heterogeneity in the historical data, each additional historical animal will thus increase the informativeness of the predicted prior distribution as compared to the informativeness of the UIP. Heterogeneity introduces additional variance in the historical means and will generally lead to a less informative prior distribution than in the case, where the prior distribution is obtained as a prediction that assumes full equality of the means in the historical and the new data. Hence, it depends on multiple components on how informative the prediction for the prior in the new experiment will be. As a result, no general statements can be made on how many animals can be saved by Bayesian predictive approaches from historical data models, although with adequate historical data, it is likely that there will be a reduction over the sample size estimates from frequentist approaches. This stresses the importance of design simulations, as performed in this thesis.

Sensitivity to the prior distribution for the residual standard deviation

In the analyzed examples, the calculated sample sizes with a partly Bayesian analysis (with a fixed residual SD as defined above in Section 4.1.2) were the same as those with from a completely Bayesian analysis (where the residual SD was assigned a prior distribution). This was true for those analyses, where the fixed SD was set equal to the (true) value, that was used for simulating the new experiment's data. Additional scenarios were examined, where the SD in the data differed from the fixed value, that was set for the partly Bayesian analysis. To model such deviation in a completely Bayesian design, the SD in the data was set to deviate from the expected value of its prior distribution (instead of a deviation from a fixed value). In those cases,

where the true SD was higher than the fixed value of its prior distribution in a partly Bayesian analysis and higher than the mean and median of its prior in a completely Bayesian analysis, the posterior distributions of the group means in the partly Bayesian analysis were too narrow. This led to an increase in the type I error over the analysis in a completely Bayesian analysis, suggesting that the completely Bayesian analysis may be preferable.

Instead of setting the fixed SD in the partly Bayesian analysis equal to the median of the predicted SD, it could be set equal to the empirical SD from the new experiment. However, due to the small sample sizes, the estimates in the data of the new experiment are not very reliable. As a consequence, in an investigation of design simulations with this alternative estimate, a lower power and higher error rates were obtained than with the estimates, that were derived from a greater data set of historical data (under the assumption that the prediction model reflects the relation of the historical and the new data correctly).

4.2.2 Results from the investigation of additional design quality criteria

As discussed in Section 4.1.2, additional design quality criteria were examined to allow for a better judgment of the benefit of using Bayesian predictive approaches over Bayesian approaches with uninformative prior distributions and over frequentist methods for the design and analysis of new animal experiment. Regarding the average width of the credible intervals, the width could be reduced by using the informative prior distributions in a Bayesian predictive approach over the widths of the frequentist confidence intervals and the widths of the Bayesian credible intervals with a less informative UIP. This decreased width is also in line with the increased power in the former approach. Assuming the distribution model, that underlies the confidence or credible interval, is correctly specified and coincides with the distribution model of the observed data, it has been shown that, when the prior variance for the tested parameter becomes sufficiently large or when the sample size becomes sufficiently large, then the limits of the confidence and credible intervals are approximately the same [59, 213]. Therefore, the potential of the Bayesian analysis is expected to be greater in cases with small sample sizes and with informative prior distributions.

On the downside of the Bayesian predictive approach, the mean of the Bayesian posterior distributions of the effect underestimates true effects greater than zero on average. This is because both prior distributions for the mean and the experimental group were centered at the mean of the predictive distribution, which was estimated from the historical data. Since the posterior distribution is a compromise between the prior and the data, the posterior treatment group mean gets pulled towards the prior mean, as compared to the frequentist estimate from the data only. As a result, estimates from the posterior treatment underestimate the true mean, when the data

estimate is greater than the prior mean (prior-data conflict). On the other hand, the Bayesian inference reduces the influence of single extreme measurements that have a large effect on the mean estimate in a frequentist analysis, as reflected in a decreased MSE and exaggeration factor over the frequentist analysis. Moreover, even with a low calculated sample size of three or four animals in the control and ten animals in the experimental group, the coverage probability of the Bayesian credible intervals was still very high (95% or higher), indicating an adequate representation of posterior uncertainty. As discussed in Section 4.1.2 in the context of mixture prior distributions with vague components, choosing a prior distribution with a larger variance will increase the impact of the data in the posterior estimates and will make them more robust for prior-data conflicts. On the other hand, the Bayes factor is very sensitive to the variance in the prior distributions and with more a priori uncertainty, it will become harder to reject the null hypothesis. As an alternative to adjusting the variance, the prior for the treatment group may be altered by assigning it a mean parameter that differs from the control mean. However, this would mean to a priori favor a decision for the alternative hypothesis and is generally not recommended [87].

4.2.3 Results from the investigated prior-data conflicts

As discussed in Section 4.1.2, three types of prior-data conflicts have been investigated in this thesis, in terms of deviations of the true mean or residual SD in the data from the expected value or the medians of the corresponding prior distributions. As discussed before, the mean parameter of the prior distribution for the experimental group mean was conservatively set equal to the mean in the control group, implying an a priori expectation of no difference in the group means. The intention behind this conservative prior mean difference of zero was, that then, a posterior non-zero mean difference may only be caused by evidence from the new experiment's data, but not by delusive a priori expectations. As discussed in the previous section, a result for the posterior distribution was, that the posterior mean in the experimental group was pulled towards the control mean, as compared to the empirical mean in the data. Concerning the first kind of prior-data conflicts (as the means of the new experiment's data deviating from the expected value of their prior distributions), this led to an underestimation of the treatment effects. For example, the design simulations of the fictive example showed that, with a prior mean of 0.13 for both groups, a true control mean of zero, and a true treatment mean of 1.5, the posterior mean in the control group would be overestimated (since the prior pulls it to higher values than the true mean of zero) and the mean of the treatment group would be underestimated (since it the prior pulls it to lower values than the true mean of 1.5).

The relative impact of the prior distribution (in relation to the impact of the data) on the posterior distribution was larger in the control group than in the experimental group, since the prior for the control group's mean was chosen to be more informative. However, since the (average) difference in the expected value of the prior mean and (estimated) data mean was larger in the experimental than in the treatment group, the bias in the posterior mean was more notable in the experimental group. Still, pulling the posterior towards the historical control group means may as well be a desirable effect that prevents researchers from claiming false significance and publishing exaggerated effects. This holds true especially in the situation of translational animal experiments, where the sample size in the new experiment is small, whereas the prior was based on a larger sample size of historical control data, and when the experimental researchers don't want to bias their conclusions from the new experiment by a priori expectations, that are unfounded.

In the fictive experiment, in both the context of planning the new experiment based on simulations from the biased historical mean and SD parameter estimates and the context of analyzing the new experiment with the true parameters, the experimental group was simulated to differ by the control group by an absolute difference of 1.5. Hence, the difference between the true experimental mean and its prior was modeled in both the biased planning data set and the true analysis data set. As a consequence, the mean difference in the simulated data did not impact the power. However, the overestimation of the true mean in the control group by its prior led to a lower power in the new experiment than the power in the situation where the prior mean in the control group would have corresponded to the true value of zero. This is because the posterior control mean gets more similar to the experimental mean than it would have been the case with a control prior mean of zero. Nonetheless, the power with the biased prior distributions and the biased starting point for the sample size calculation was still 84% with the data that were simulated from true parameters. For the investigation of type I errors, the true mean in the experimental group was simulated to be zero. In this situation and with biased prior means (as estimates from the historical data), the choice of a more informative prior for the control than for the experimental group can increase the type I error rate, although the means of the prior distributions in the two groups were set to be identical. This is because the posterior mean in the control group will be pulled more towards the biased value of the prior than in the case of the experimental group. However, in the analyzed example, the type I error was only 1% for both the biased planning and the true analysis data set. Thus, in this example, the type I error rate would still be below 5% in the Bayesian analysis and the power above 80%. Hence, the Bayesian sample size estimate did not imply an underpowered design, in contrast to the frequentist sample size estimate, while the Bayesian approach still had the advantage of a lower control group sample size.

Regarding the second kind of prior-data conflicts (deviation of the mean and median of the prior distribution for the predicted control SD from the new experiment's true parameter), this was investigated in the fictive example of multi-experiment data. There, the estimates of the overall mean and residual SD, that were obtained from the historical data, deviated from the true values, that were used for simulating the data. By summarizing the historical data of the simulated five small animal experiments in a meta-analysis, the deviations of the single experiments could be ameliorated. Still, the expected value of the overall mean differed by 0.13 from the true control mean. In addition, the estimate of the SD was slightly underestimated with a median of 0.98 in the posterior distribution, while the true SD was one. The estimates from the Bayesian meta-analysis were used to get a starting point for sample size reduction with the frequentist two-sample t-test. A minimal clinically relevant treatment effect of 1.5 was used and a SD, that was obtained from the posterior distribution of the Bayesian meta-analysis. As discussed above in the section on the sample size results, this led to a starting point for the Bayesian sample size approach of eight animals per group. In contrast, a sample size of nine animals would have been obtained with the frequentist two-sample t-test for the true parameter values (that were used for simulating the historical data). Hence, the starting point for sample size estimation was slightly too low compared to the true value. This would imply an underpowered design with the frequentist t-test. Still, as stated above, the benefit of planning the new experiments based on the estimates from a meta-analysis of multiple historical experiments over a sample size planning based on the estimates from a single historical experiment becomes clear, since the estimated sample size for the new experiment with the former meta-analysis approach was, on average, notably closer to the true sample size, than when the latter approach.

Regarding the third kind of prior-data conflicts (deviations of the means or medians of the prior distribution for both the predicted control mean and SD from the new experiment's true parameters), the predicted prior distribution for the residual SD was modified to have an expected value of either 1.5 or 0.5 times the true SD of one, while, for the mean parameter, the same predicted prior distribution (estimated from the historical data model) as before was used. The larger expected value of the SD prior of 1.5 did not impact the power and type I error much, compared to the power and type I error, that were obtained with a SD whose expected was approximately equal to the true SD of the new experiment's simulated data. The smaller expected value of the SD prior of 0.5 led to an increase in both the power to 93% (for the cases where the true mean difference was 1.5) and an increase of the type I error to 16% (for the cases where the true mean difference was zero). Still, modeling the residual SD as a random variable instead of using a fixed estimate from the historical data (or the new data) helped to decrease the type I error rate in such cases. This is because the model with the SD as a random variable led to larger uncertainty in the estimation of the mean difference in terms of wider credible intervals. False claims

of significance may as well be a problem in frequentist analysis, especially when the experiment is underpowered, as it is common for animal experiments [8].

The impact of the prior may be unwanted in cases of apparent prior-data conflict when the frequentist empirical estimates from the data have only a small bias (i.e. when the estimates have a small variance and expected values that correspond to the true parameter values). However, as argued above, in those cases, where the frequentist estimates are impacted by outliers and are biased as compared to the true parameter values, it may as well be that the estimate from a larger sample of historical data is better. In this case, the impact of the prior distribution may be wanted, since, depending on whether the residual SDs and the control group mean are over- or underestimated by the data, the prior distributions from the historical data can decrease the risk of false significant claims and of exaggeration of effects, or they can increase the power, which are desirable properties in light of reproducible research.

4.3 Conclusions

Regarding the main research question, it can be concluded, that a systematic approach to estimating the sample size for a new animal experiment using Bayesian predictive prior distributions from historical individual-level animal data could be established and that the proposed approach has a high potential for the planning and analysis of translational animal experiments. In standard frequentist approaches, the sample size for a new experiment is often calculated with the parameter estimates from a single small historical experiment without taking into account the uncertainty in the estimation. The simulation of such small experiments in a fictive example showed, that the low sample sizes of the historical experiments, and the omission of the uncertainty about the parameters in the frequentist approach, result in unreliable parameter estimates and in turn to unreliable sample size estimates for the new experiment. This demonstrates the limited value of classical sample size determination and emphasizes the need for alternative methods [7]. The simulations showed that a better planning foundation can be obtained by combining the information from the historical data and deriving predictions for the new experiment by meta-analytic methods. The predictions should be made in consideration of the relationship between the historical and the new experiment, which extends the classical assumption of identical parameters. Modeling the historical data on an individual-animal level instead of using aggregated summary data (as arithmetic means and empirical SDs) allows for more complex models and the representation of subgroup effects, like differences due to the impacts of the animals' age or sex, and thus provides a higher information content for the new experiment than meta-analyses of summary data. In addition, the flexible regression framework and the employed software implementation

allow for the application of the approach to a large variety of distribution models, including the representation of unequal variances in the treatment groups (heteroscedasticity). The design simulations showed that the use of Bayesian predictive prior distributions in the analysis of the new experiment can decrease the chance of false claims of significant treatment effects by decreasing the impact of single extreme measurements and can lead to less exaggerated effects compared to standard frequentist approaches.

With the design simulation approach, a flexible framework for determining the new experiment's sample size is provided by simulating hypothetical data for the new experiment and repeatedly evaluating design criteria like the statistical power and the type I error rate of a hypothesis test. The proposed simulation procedure can serve both as a basis for deciding on a final sample and as a sensitivity analysis to different model assumptions. The design simulations showed, that if the prediction model reflects the actual relationship between the historical and the new data, then the inclusion of a priori information from historical control animals in the planning and analysis of a new experiment in a Bayesian predictive approach allows for a reduction of the overall sample size for the new animal experiment compared to the sample size obtained with a standard frequentist approach while maintaining a target power of at least 80% and a type I error rate below 5% of the hypothesis test.

In the frequentist analysis of homoscedastic data, the greatest power is generally achieved in balanced designs with an equal allocation rate of the overall sample size to the treatment groups. In contrast, an allocation of a greater proportion of the overall sample size to the experimental than to the control group is preferable in a Bayesian analysis in a situation where informative prior distributions can be obtained from the historical control data. This is because the amount of new evidence, that is necessary to reach a certain statistical power (like 80%), can be lowered by including the information from the available control data. For experimental groups, in which new treatments are investigated, typically no specific prior knowledge about the new experimental treatment is available. Thus, a greater amount of new evidence is necessary for this group, as reflected in the allocation of a larger proportion of the overall sample size to the experimental group than to the control group. In contrast, in the frequentist analysis or in the Bayesian analysis with non-informative prior distributions that don't use historical data, it is necessary to collect the same amount of information for both treatment groups in the new experiment. In theory, predictive prior distributions could be obtained for the experimental group as well. However, the historical and new experimental animals are typically less similar than the control animals [8]. As a consequence, the Bayesian predictive approach is expected to have a greater potential in reducing the control group sample size than in reducing the sample size in the experimental group.

Regarding the secondary research questions, in the analyzed examples, the sample size decisions with a hypothesis test based on whether the Bayes factor exceeded a threshold of three (as moderate evidence for the alternative hypothesis [56]) were the same as the sample size decisions with a hypothesis test based on whether the 95% equal-tailed interval (ETI) included the hypothesized parameter value under the null hypothesis of equal treatment group means. In this situation, researchers may prefer the ETI hypothesis test for its simplicity. However, the Bayes factor hypothesis test has other interpretational advantages [54] making it the recommended default choice for Bayesian hypothesis testing [108, 109, 146]. Furthermore, in the analyzed scenarios, the sample size decisions with a partly Bayesian analysis with a fixed residual SD were the same as those with a completely Bayesian analysis where the residual SD was modeled as a random variable. This was true for those scenarios, where the fixed SD was set equal to the (true) value, that was used for simulating the new experiment's data. However, the investigation of scenarios where the information from the predicted prior distribution is in conflict with the information in the new data (prior-data conflicts), showed that the type I error in a partly Bayesian analysis with fixed SD can be higher than the type I error rate in a completely Bayesian analysis. This suggests that the completely Bayesian analysis may be preferred over a partly Bayesian analysis since it better reflects the uncertainty in the predicted parameters. The influence of the prior distribution of the mean parameter on the sample size decision was examined by comparing the results with an informative predicted prior distribution to the results with a unit information prior (UIP), which is, by definition, as informative as one single animal from the new experiment. The required sample size with the UIP was larger than the required sample size with the informative predicted prior distribution, stressing the importance of the historical data.

To obtain useful sample size estimates, it is important to consider statistical goals that are in line with the actual research purpose. In the simulated data, additional design quality criteria, like the widths of credible intervals, the average errors of point estimates, or the average exaggeration factors of significant results, were examined, which led to a better understanding of the experimental designs and thus enhances reproducible research. Instead of using the Bayesian predictive approach to reduce the sample size in future experiments, the additional information implied by a Bayesian predictive approach can also be used to keep the frequentist sample size at the same value, but instead increase the power of the Bayesian test (and thus, by definition, the probability of detecting truly present effects) in the new experiment over the power of the frequentist test, or to allow for the detection of lower effect sizes than the ones that can be detected with standard frequentist tests.

In summary, it can be concluded that the Bayesian predictive approach to the planning and analysis of translational animal experiments outperforms classical frequentist approaches, firstly,

by allowing for a reduction of a new experiment's sample size while reflecting the uncertainty about the parameters in the new data. This is enabled by the inclusion of the predictions from the historical data as prior distribution in a Bayesian analysis of the new data. Secondly, the approach outperforms classical frequentist approaches by allowing for more precise and realistic parameter estimates for the new experiment. This is enabled by combining historical data from several experiments in a common model and reflecting sources of heterogeneity that extend the assumption of identical parameters in the historical and the new data by using meta-analytic methods in the Bayesian predictions and analyses. With these features, the proposed Bayesian predictive approach to sample size determination in translational animal experiments provides both aspects that contribute to the reduction aspect of the 3R concept [2] and also aspects that help to ameliorate the reproducibility issues in animal research [8].

4.4 Outlook

To get the greatest possible benefit from historical data, a collection in a common database with a uniform vocabulary, standards, and formats is desirable. Such databases are initiated by several institutions. In this thesis, data from the Mouse Phenome Database (MPD) have been considered, which was developed by the Mouse Genome Informatics (MGI) and is headquartered by the Jackson Laboratory [58]. The MPD receives worldwide contributions of phenotyping outcomes and disease-related traits in both control and treated mice. The MGI also integrates mutant data from the International Mouse Phenotyping Consortium (IMPC), which is an international effort for characterizing common phenotypes of knockout mouse strains [159]. The IMPC consists of over twenty international research institutes, including the Jackson Laboratory in the United States of America and the Helmholtz Zentrum München in Germany. Data in the IMPC have undergone quality checks (for example checking outliers) and are endowed with both the Mammalian Phenotype ontology by the MGI, in compliance with the MPD, and the Human Phenotpye ontology, which facilitates the interpretation of the data for translational and clinical researchers. Further efforts for data collection from different research areas include the CASIMIR program (Coordination and Sustainability of International Mouse Informatics Resources) [216] for the coordination and integration of databases, the tumor database RITA [217] of the Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM) in Hannover and the projects eTOX [218] and eTRANSAFE [219] of the Innovative Medicine Initiatives for data exchange in toxicological and translational experiments on the safety of drugs. A challenge that has existed so far is that many institutions cannot or do not want to make their data publicly available. Although the publication of animal data may be easier than human data, e.g., because animal data do not have to be anonymized,

one existing difficulty is the protection against intellectual theft. As a solution, concepts are being developed to partially share data (*partial data sharing*) [219].

With the emergence and development of phenotypic databases and since the amount of historical data is increasing with every conducted animal experiment, it seems wasteful to not use the available historical information. Hence, Bayesian predictive approaches, which (meta-analytically) combine and re-use historical information, are likely to become even more relevant in the future. Although Bayesian analyses are typically more complex than frequentist analyses, tools for making them more accessible and easier to apply are continuously developed (e.g., [8, 65, 105, 106, 220]). This will likely extend the user community to more applied researchers from the medical field. Finally, there have been calls to expand statistical consultation capacities and to develop user guides and improved biometrical forms [6]. These efforts are likely to facilitate transitions from frequentist to Bayesian approaches for translational researchers with little statistical background.

5 Summary

The reduction of sample size in animal experiments is an essential contribution to animal welfare. However, the problem is that, generally, with a smaller sample size, the precision of estimates decreases, making it harder to answer the medical question of interest. This thesis introduces a systematic approach to sample size estimation for a new animal experiment by using Bayesian predictive prior distributions from historical individual-level animal data. It has been investigated if this approach can allow for a reduction of sample size compared to the sample size, which would be required in a classical frequentist approach to maintain a certain statistical power and type I error rate of the corresponding hypothesis tests. The idea is to gather data from historical experiments, that address the same pathological phenotype with possibly different experimental treatments, but similar control groups, and to tackle the problems, that are associated with the frequentist planning and analysis of single small translational animal experiments, by using:

(a) Meta-analytic methods to combine historical control data and make predictions for a new animal experiment, with the aim of obtaining more reliable estimates and a better planning basis while reflecting the relationship between the historical and the new experiments

(b) Bayesian methods to incorporate the predictions from the model of the historical data as prior distributions into the analysis of the new experiment, with the aim of reducing the sample size in the new experiment while reflecting the uncertainty in the predictions

Various ways for making predictions for the mean and residual standard deviation of hypothetical observations in the new experiment have been discussed in light of the assumed relationship between the historical and the new experiment's data. Random variation of the experiment-specific parameters around an overall parameter can be modeled as heterogeneity, relaxing the assumption of identical parameters in all experiments. Fitting models to historical individual-level animal data instead of aggregated summary data allows for more informative predictions and more flexibility in modeling subgroup effects. Predictions for the new experiment were obtained from the posterior distributions of the Bayesian model of the historical data. Design simulations were used to investigate whether the Bayesian predictive approach allows for a reduction of the control group sample size over the estimate obtained with a frequentist two-sample t-test or Welch test at a two-sided significance level of 5% and a target power of at least 80%. The

reduction of sample size was only considered for the control but not for the experimental group, since no historical data on the particular experimental treatment were available and the historical data of other treatments were assumed to be less similar than the control data. The Bayesian hypothesis tests were based on whether the Bayes factor exceeded a threshold of moderate evidence for the alternative hypothesis or whether the 95% equal-tailed interval (ETI) included the parameter value of the null hypothesis. The proposed approaches were illustrated in a real-world animal experiment examining the knockout effect of the C5aR1 in osteoclasts and osteoblasts on bone quality as measured by the two experimental outcomes bending stiffness and relative bone volume (C5aR1 example). Internal control data were provided by the experiment's researcher and predictions were made under the assumption, that the new control data represent additional measurements to the historical control data. Further external data from the Mouse Phenome Database (MPD) for the same experimental could not be used since the definition of the control group in the MPD data differed from the control definition of the internal data. In addition to the C5aR1 example, a fictive example with historical data from five small historical experiments was investigated, where predictions were made under the weaker assumption that the means in the historical and the new control animals are similar but not identical.

For the first outcome in the C5aR1 example, the estimated control group sample size of ten animals with the frequentist two-sample t-test could be reduced to four animals and for the second outcome, it could be reduced to three animals. For the fictive example, the limitations of the criticized approach of calculating the new experiment's sample size with the parameter estimates from a single small experiment were demonstrated. In the criticized approach, the largest calculated sample size was 344 animals per treatment group and the smallest sample size was seven animals, whereas the sample size with the true parameters was nine animals per group. A better sample size estimate could be obtained by combining the historical data in a meta-analysis. Then, using the Bayesian predictive approach allowed for a reduction of the control group sample size to three animals. The sample size decisions were the same with either the Bayes factor or the ETI based Bayesian hypothesis test. However, modeling the standard deviation as a random variable instead of a fixed value was advantageous regarding the type I error rate in case of conflicts between the information from the data and the prior distributions. The design simulations suggested that the proposed Bayesian predictive approach allows for narrower credible intervals with a likewise higher probability of covering the true effect and less exaggerated effect estimates in cases of significant findings than classical frequentist approaches, which helps to increase the reproducibility of the new experiment's results. In conclusion, Bayesian predictive approaches have a high potential for the planning and analysis of translational animal experiments, and, given current efforts to collect phenotypic animal data more systematically, an even greater potential of these methods is expected in the future.

6 References

[1] T. Unseld. Designing translational animal experiments by Bayesian MAP approaches. *arXiv* 2301.05572 (2023). URL: https://arxiv.org/abs/2301.05572.

[2] W. Russell and R. Burch. The principles of humane experimental technique. London: Methuen, 1959.

[3] Bundesministerium der Justitz. Verordnung zum Schutz von zu Versuchszwecken oder zu anderen wissenschaftlichen Zwecken verwendeten Tieren (Tierschutz-Versuchstierverordnung). 2010. URL: https://www.gesetze-im-internet.de/tierschversv/BJNR312600013.html (visited on 05/08/2023).

[4] Bundesministerium der Justitz. Deutsches Tierschutzgesetz. 1972. URL: https://www.gesetze-im-internet.de/tierschg/ (visited on 05/08/2023).

[5] Die Bundesregierung. Allgemeine Verwaltungsvorschrift zur Durchführung des Tierschutzgesetzes. 2000. URL: https://www.verwaltungsvorschriften-im-internet.de/bsvwvbund_09022000_32135220006.htm (visited on 05/08/2023).

[6] S. K. Piper, D. Zocholl, U. Toelch, R. Roehle, A. Stroux, J. Hoessler, A. Zinke, and F. Konietschke. Statistical review of animal trials – a guideline. *Biometrical Journal* 65.2 (2023), 2200061.

[7] B. Mayer and R. Muche. Die limitierte Aussagekraft formaler Fallzahlplanung im Rahmen von Tierversuchen der medizinischen Grundlagenforschung. *Tierärztliche Praxis Ausgabe K: Kleintiere/Heimtiere* 41.06 (2013), 367–374.

[8] V. Bonapersona, H. Hoijtink, R. Sarabdjitsingh, and M. Joëls. Increasing the statistical power of animal experiments with historical control data. *Nature Neuroscience* 24.4 (2021), 470–477.

[9] J. C. De Winter. Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation* 18.10 (2013). URL: https://scholarworks.umass.edu/pare/vol18/iss1/10.

[10] M. J. Campbell, S. A. Julious, and D. G. Altman. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Medical Journal* 311.7013 (1995), 1145–1148.

[11] S. Janušonis. Comparing two small samples with an unstable, treatment-independent baseline. *Journal of Neuroscience Methods* 179.2 (2009), 173–178.

[12] A. Gelman and J. Carlin. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9.6 (2014), 641–651.

[13] M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour* 1.1 (2017), 1–9.

[14] B. Voelkl and H. Würbel. Reproducibility crisis: are we ignoring reaction norms? *Trends in Pharmacological Sciences* 37.7 (2016), 509–510.

[15] M. Macleod and S. Mohan. Reproducibility and rigor in animal-based research. *Institute for Laboratory Animal Research Journal* 60.1 (2019), 17–23.

[16] U. Dirnagl. Resolving the tension between exploration and confirmation in preclinical biomedical research. *Good Research Practice in Non-Colinical Pharmacology and Biomedicine* (2019), 71–79.

[17] B. Mayer, A. Allgöwer, and R. Muche. Notwendige Standards der biometrischen Fallzahlplanung von Tierversuchen in der präklinischen Forschung im Sinne der 3R. *Berliner und Münchener Tierärztliche Wochenschrift* 131 (2018), 272–278.

[18] J. P. Ioannidis. Why most published research findings are false. *PLoS medicine* 2.8 (2005), e124.

[19] F. S. Collins and L. A. Tabak. Policy: NIH plans to enhance reproducibility. *Nature* 505.7485 (2014), 612.

[20] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe. The economics of reproducibility in preclinical research. *PLoS Biol* 13.6 (2015), e1002165.

[21] C. G. Begley and J. P. Ioannidis. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research* 116.1 (2015), 116–126.

[22] R. L. Jilka. The road to reproducibility in animal research. *Journal of Bone and Mineral Research* 31.7 (2016), 1317–1319.

[23] B. Voelkl, N. S. Altman, A. Forsman, W. Forstmeier, J. Gurevitch, I. Jaric, N. A. Karp, M. J. Kas, H. Schielzeth, T. Van de Casteele, and H. Würbel. Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience* 21.7 (2020), 384–393.

[24] N. I. Drude, L. Martinez Gamboa, M. Danziger, U. Dirnagl, and U. Toelch. Science Forum: improving preclinical studies through replications. *eLife* 10 (2021), e62101.

[25] S. Novick and T. Zhang. Mean comparisons and power calculations to ensure reproducibility in preclinical drug discovery. *Statistics in Medicine* 40.6 (2021), 1414–1428.

[26] M. R. Macleod, S. Ebrahim, and I. Roberts. Surveying the literature from animal experiments: systematic review and meta-analysis are important contributions. *British Medical Journal* 331.7508 (2005), 110.

[27] C. Hooijmans, M Rovers, R. de Vries, M Leenaars, and M Ritskes-Hoitinga. An initiative to facilitate well-informed decision-making in laboratory animal research: report of the First International Symposium on Systematic Reviews in Laboratory Animal Science. *Laboratory Animals* 46.4 (2012), 356–357.

[28] E. S. Sena, G. L. Currie, S. K. McCann, M. R. Macleod, and D. W. Howells. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *Journal of Cerebral Blood Flow and Metabolism* 34.5 (2014), 737–742.

[29] R. Greek and A. Menache. Systematic reviews of animal models: methodology versus epistemology. *International Journal of Medical Sciences* 10.3 (2013), 206–221.

[30] R. B. de Vries, K. E. Wever, M. T. Avey, M. L. Stephens, E. S. Sena, and M. Leenaars. The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *Institute for Laboratory Animal Research Journal* 55.3 (2014), 427–437.

[31] C. R. Hooijmans, J. IntHout, M. Ritskes-Hoitinga, and M. M. Rovers. Meta-analyses of animal studies: an introduction of a valuable instrument to further improve healthcare. *Institute for Laboratory Animal Research Journal* 55.3 (2014), 418–426.

[32] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. Bayesian approaches to clinical trials and health-care evaluation. Vol. 13. Chichester: John Wiley & Sons, 2004.

[33] R. Walley, J. Sherington, J. Rastrick, E. Detrait, E. Hanon, and G. Watt. Using Bayesian analysis in repeated preclinical in vivo studies for a more effective use of animals. *Pharmaceutical Statistics* 15.3 (2016), 277–285.

[34] M. Kramer and E. Font. Reducing sample size in experiments with animals: historical controls and related strategies. *Biological Reviews* 92.1 (2017), 431–445.

[35] H. Yang and S. Novick. Bayesian analysis with R for drug development: concepts, algorithms, and case studies. Boca Raton: Chapman and Hall/CRC, 2019.

[36] A. Gelman and M. Vákár. Slamming the sham: A Bayesian model for adaptive adjustment with noisy control data. *Statistics in Medicine* 40.15 (2021), 3403–3424.

[37] T. Friede, C. Röver, S. Wandel, and B. Neuenschwander. Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods* 8.1 (2017), 79–91.

[38] N. Kallioinen, T. Paananen, P.-C. Bürkner, and A. Vehtari. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *arXiv* 2107.14054 (2023). URL: https://arxiv.org/abs/2107.14054v4.

[39] B. Neuenschwander, G. Capkun-Niggli, M. Branson, and D. J. Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials* 7.1 (2010), 5–18.

[40] N. W. Galwey. Supplementation of a clinical trial by historical control data: is the prospect of dynamic borrowing an illusion? *Statistics in Medicine* 36.6 (2017), 899–916.

[41] J. M. Bernardo. The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences* 4 (1996), 111–122.

[42] E. M. Pullenayegum. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine* 30.26 (2011), 3082–3094.

[43] K. M. Rhodes, R. M. Turner, and J. P. Higgins. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology* 68.1 (2015), 52–60.

[44] R. M. Turner, D. Jackson, Y. Wei, S. G. Thompson, and J. P. Higgins. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine* 34.6 (2015), 984–998.

[45] F. Bartoš, Q. F. Gronau, B. Timmers, W. M. Otte, A. Ly, and E.-J. Wagenmakers. Bayesian model-averaged meta-analysis in medicine. *Statistics in Medicine* 40.30 (2021), 6743–6761.

[46] H. Schmidli, B. Neuenschwander, and T. Friede. Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials. *Computational Statistics & Data Analysis* 113 (2017), 100–110.

[47] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv* 2011.01808 (2020). URL: https://arxiv.org/abs/2011.01808.

[48] L. A. Stewart and J. F. Tierney. To IPD or not to IPD?: advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions* 25.1 (2002), 76–97.

[49] S. Wandel, B. Neuenschwander, C. Röver, and T. Friede. Using phase II data for the analysis of phase III studies: an application in rare diseases. *Clinical Trials* 14.3 (2017), 277–285.

[50] J. G. Ibrahim and M.-H. Chen. Power Prior distributions for regression models. *Statistical Science* 15.1 (2000), 46–60.

[51] E.-J. Wagenmakers, R. D. Morey, and M. D. Lee. Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science* 25.3 (2016), 169–176.

[52] F. D. Schönbrodt and E.-J. Wagenmakers. Bayes factor design analysis: planning for compelling evidence. *Psychonomic Bulletin and Review* 25.1 (2018), 128–142.

[53] R. Kelter. Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests. *WIREs Computational Statistics* 13.6 (2021), e1523.

[54] E.-J. Wagenmakers, M. D. Lee, J. N. Rouder, and R. D. Morey. The principle of predictive irrelevance, or why intervals should not be used for model comparison featuring a point null hypothesis. *PsyArXiv* (2019). URL: psyarxiv.com/rqnu5.

[55] H. Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 31. 2. Campridge: Cambridge University Press, 1935, 203–222.

[56] A. M. Stefan, Q. F. Gronau, F. D. Schönbrodt, and E.-J. Wagenmakers. A tutorial on Bayes Factor design analysis using an informed prior. *Behavior Research Methods* 51.3 (2019), 1042–1058.

[57] S. Vasishth, H. Yadav, D. J. Schad, and B. Nicenboim. Sample size determination for Bayesian hierarchical models commonly used in psycholinguistics. *Computational Brain and Behavior* (2022), 1–25.

[58] M. A. Bogue, S. C. Grubb, D. O. Walton, V. M. Philip, G. Kolishovski, T. Stearns, M. H. Dunn, D. A. Skelly, B. Kadakkuzha, G. TeHennepe, G. Kunde-Ramamoorthy, and E. J. Chesler. Mouse Phenome Database: an integrative database and analysis suite for curated empirical phenotype data from laboratory mice. *Nucleic Acids Research* 46.D1 (2018), D843–D850.

[59] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Bayesian data analysis. 3rd ed. New York: Chapman and Hall/CRC, 2013.

[60] R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90.431 (1995), 928–934.

[61] L. Al Labadi and M. Evans. Optimal robustness results for relative belief inferences and the relationship to prior-data conflict. *Bayesian Analysis* 12.3 (2017), 705–728.

[62] M. Evans and H. Moshonov. Checking for prior-data conflict. *Bayesian Analysis* 4.1 (2006), 893–914.

[63] M. Reimherr, X.-L. Meng, and D. L. Nicolae. Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.3 (2021), 413–437.

[64] T. Mutsvari, D. Tytgat, and R. Walley. Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics* 15.1 (2016), 28–36.

[65] P.-C. Bürkner. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software* 100.5 (2021), 1–54.

[66] R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, 2022. URL: https://www.R-project.org/.

[67] C. Kilkenny and D. G. Altman. Improving bioscience research reporting: ARRIVE-ing at a solution. *Laboratory Animals* 44.4 (2010), 377–378.

[68] N. P. du Sert, V. Hurst, A. Ahluwalia, S. Alam, M. T. Avey, M. Baker, W. J. Browne, A. Clark, I. C. Cuthill, U. Dirnagl, M. Emerson, P. Garner, S. T. Holgate, D. W. Howells, N. A. Karp, S. E. Lazic, K. Lidster, C. J. MacCallum, M. Macleod, E. J. Pearl, O. H. Petersen, F. Rawle, P. Reynolds, K. Rooney, E. S. Sena, S. D. Silberberg, T. Steckler, and H. Würbel. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology* 18.7 (2020), 1–12.

[69] University Hospital of Neurology Ulm. Hosptial Report 2020. URL: https://www.uniklinik-ulm.de/fileadmin/default/Kliniken/Neurologie/Klinikberichte/Klinikbericht_2020_HP.pdf (visited on 11/24/2022).

[70] University Hospital of Neurology Ulm. Hosptial Report 2022. URL: https://www.uniklinik-ulm.de/fileadmin/default/Kliniken/Neurologie/Klinikberichte/Klinikbericht_2022_Inhalt_S1-272_mit_Cover_HP.pdf (visited on 11/24/2022).

[71] E. L. Lehmann and J. P. Romano. Testing statistical hypotheses. 3rd ed. New York: Springer, 2005.

[72] B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29.3/4 (1938), 350–362.

[73] F. E. Satterthwaite. Synthesis of variance. *Psychometrika* 6.5 (1941), 309–316.

[74] M. Kieser. Methods and applications of sample size calculation and recalculation in clinical trials. Cham: Springer, 2020.

[75] M. Happ, A. C. Bathke, and E. Brunner. Optimal sample size planning for the Wilcoxon-Mann-Whitney test. *Statistics in Medicine* 38.3 (2019), 363–375.

[76] C. T. Ekstrøm. Miscellaneous Esoteric Statistical Scripts. R package version 0.5.12. 2023. URL: https://CRAN.R-project.org/package=MESS (visited on 11/14/2023).

[77] A. Heiss. A guide to correctly calculating posterior predictions and average marginal effects with multilievel Bayesian models. 2021. URL: https://www.andrewheiss.com/blog/2021/11/10/ame-bayes-re-guide/ (visited on 06/09/2023).

[78] P. Bürkner. Estimating distributional models with brms. 2020. URL: https://cran.r-project.org/web/packages/brms/vignettes/brms_distreg.html (visited on 06/09/2023).

[79] D. J. Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII — 1983*. Berlin: Springer Berlin Heidelberg, 1985, 1–198.

[80] P. D. Johnson and D. G. Besselsen. Practical aspects of experimental design in animal research. *Institute for Laboratory Animal Research Journal* 43.4 (2002), 202–206.

[81] L. V. Hedges and I. Olkin. Statistical methods for meta-analysis. San Diego: Academic Press, 1985.

[82] J. Hartung and G. Knapp. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* 20.24 (2001), 3875–3889.

[83] J. P. Higgins, J. Thomas, J Chandler, T. Cumpston M. Li, M. Page, and W. V.A. Cochrane handbook for systematic reviews of interventions. Chichester: John Wiley & Sons, 2019.

[84] H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 70.4 (2014), 1023–1032.

[85] B. Neuenschwander, S. Wandel, S. Roychoudhury, and S. Bailey. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics* 15.2 (2016), 123–134.

[86] C. Röver and T. Friede. Bounds for the weight of external data in shrinkage estimation. *Biometrical Journal* 63.5 (2021), 1131–1143.

[87] D. J. Schad, M. Betancourt, and S. Vasishth. Toward a principled Bayesian workflow in cognitive science. *Psychological Methods* 26.1 (2021), 103–126.

[88] M. Betancourt. How the shape of a weakly informative prior affects inferences. 2017. URL: https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html (visited on 05/23/2021).

[89] J. W. Seaman III, J. W. Seaman Jr, and J. D. Stamey. Hidden dangers of specifying noninformative priors. *The American Statistician* 66.2 (2012), 77–84.

[90] N. P. Lemoine. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos* 128.7 (2019), 912–928.

[91] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1.3 (2006), 515–534.

[92] C. Röver, R. Bender, S. Dias, C. H. Schmid, H. Schmidli, S. Sturtz, S. Weber, and T. Friede. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods* 12.4 (2021), 448–474.

[93] M. H. DeGroot. Optimal statistical decisions. Hoboken: MCGraw-Hill, 1970.

[94] G. E. Box and G. C. Tiao. Bayesian inference in statistical analysis. New York: John Wiley & Sons, 2011.

[95] A. O'Hagan and J. Forster. Kendall's advanced theory of statistics. Vol. 2B. Bayesian Inference. London: Arnold, 2004.

[96] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21.6 (1953), 1087–1092.

[97] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57.1 (1970), 97–109.

[98] Novartis Pharma AG, S. Weber, B. Neuenschwander, H. Schmidli, B. Magnusson, Y. Li, S. Roychoudhury, Trustees of Columbia University. R Bayesian Evidence Synthesis Tools. 2023. URL: https://cloud.r-project.org/web/packages/RBesT/RBesT.pdf (visited on 10/10/2023).

[99] G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician* 39.2 (1985), 83–87.

[100] B. P. Carlin and T. A. Louis. Empirical Bayes: past, present and future. *Journal of the American Statistical Association* 95.452 (2000), 1286–1289.

[101] J. P. Higgins and S. G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21.11 (2002), 1539–1558.

[102] P. Mikkola, O. A. Martin, S. Chandramouli, M. Hartmann, O. A. Pla, O. Thomas, H. Pesonen, J. Corander, A. Vehtari, S. Kaski, P.-C. Bürkner, and A. Klami. Prior knowledge elicitation: the past, present, and future. *Bayesian Analysis* (2023), 1 –33.

[103] I. J. Good. Probability and the weighing of evidence. London: Charles Griffin & Company, 1950.

[104] P. Schwaferts and T. Augustin. Bayes factors can only quantify evidence w.r.t. sets of parameters, not w.r.t. (prior) distributions on the parameter. *arXiv* 2110.09871 (2021). URL: https://arxiv.org/abs/2110.09871.

[105] C. Röver. Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software* 93.6 (2020), 1–51.

[106] S. Weber, Y. Li, J. W. Seaman, T. Kakizume, and H. Schmidli. Applying meta-analytic-predictive priors with the R Bayesian Evidence Synthesis Tools. *Journal of Statistical Software* 100.19 (2021), 1–32.

[107] B. Neuenschwander, S. Weber, H. Schmidli, and A. O'Hagan. Predictively consistent prior effective sample sizes. *Biometrics* 76.2 (2020), 578–587.

[108] S. M. Lewis and A. E. Raftery. Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association* 92.438 (1997), 648–655.

[109] J. Berger. Bayes factors. *Encyclopedia of Statistical Sciences*. 2nd ed. Wiley, 2006, 378–386.

[110] L. Held and D. Sabanés Bové. Applied statistical inference. Berlin: Springer, 2014.

[111] Y. L. Mithat Gönen Wesley O Johnson and P. H. Westfall. The Bayesian two-sample t test. *The American Statistician* 59.3 (2005), 252–257.

[112] J. M. Dickey and B. Lientz. The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics* (1970), 214–226.

[113] L. Ebner, P. M. Schwaferts, and T. Augustin. Robust Bayes factor for independent two-sample comparisons under imprecise prior information. *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*. Vol. 103. Ghent: Proceedings of Machine Learning Research, 2019.

[114] D. V. Lindley. Basic ideas in Bayesian statistics. *Bayesian Statistics: A Review*. 6th ed. Philadelphia: Society for Industrial and Applied Mathematics, 1972, 30–31.

[115] E.-J. Wagenmakers, T. Lodewyckx, H. Kuriyal, and R. Grasman. Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cognitive Psychology* 60.3 (2010), 158–189.

[116] M. S. Bartlett. A comment on D.V. Lindley's statistical paradox. *Biometrika* 44.3-4 (1957), 533–534.

[117] D. V. Lindley. A statistical paradox. *Biometrika* 44.1/2 (1957), 187–192.

[118] Q. Zhou and Y. Guan. On the null distribution of Bayes factors in linear regression. *Journal of the American Statistical Association* 113.523 (2018), 1362–1371.

[119] M. D. Lee and E.-J. Wagenmakers. Bayesian cognitive modeling: a practical course. Cambridge: Cambridge University Press, 2014.

[120] H. Jeffreys. The theory of probability. Oxford: Oxford University Press, 1939.

[121] J. Elsey. Powerful sequential designs using Bayesian estimation: a power analysis tutorial using brms, the tidyverse, and furrr. *PsyArXiv* (2021). URL: osf.io/preprints/psyarxiv/kt4pz.

[122] A. Gelman and F. Tuerlinckx. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15.3 (2000), 373–390.

[123] J. Lu, Y. Qiu, and A. Deng. A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology* 72.1 (2019), 1–17.

[124] A. Gelman, J. Hill, and A. Vehtari. Design and sample size decisions. *Regression and other stories*. Cambridge: Cambridge University Press, 2020, 291–312.

[125] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14.5 (2013), 365–376.

[126] L. S. Richelson, H. W. Wahner, L. J. r. Melton, and B. L. Riggs. Relative contributions of aging and estrogen deficiency to postmenopausal bone loss. *The New England Journal of Medicine* 311.20 (1984), 1273–1275.

[127] M. N. Weitzmann and R. Pacifici. Estrogen deficiency and bone loss: an inflammatory tale. *The Journal of Clinical Investigation* 116.5 (2006), 1186–1194.

[128] F. T. Beil, F. Barvencik, M. Gebauer, S. Seitz, J. M. Rueger, A. Ignatius, P. Pogoda, T. Schinke, and M. Amling. Effects of estrogen on fracture healing in mice. *Journal of Trauma and Acute Care Surgery* 69.5 (2010).

[129] Y. Mödinger, A. Rapp, J. Pazmandi, A. Vikman, K. Holzmann, M. Haffner-Luntzer, M. Huber-Lang, and A. Ignatius. C5aR1 interacts with TLR2 in osteoblasts and stimulates the osteoclast-inducing chemokine CXCL10. *Journal of Cellular and Molecular Medicine* 22.12 (2018), 6002–6014.

[130] D. Ricklin and J. D. Lambris. Complement-targeted therapeutics. *Nature Biotechnology* 25.11 (2007), 1265–1275.

[131] A. Ignatius, P. Schoengraf, L. Kreja, A. Liedert, S. Recknagel, S. Kandert, R. E. Brenner, M. Schneider, J. D. Lambris, and M. Huber-Lang. Complement C3a and C5a modulate osteoclast formation and inflammatory response of osteoblasts in synergism with IL-1β. *Journal of Cellular Biochemistry* 112.9 (2011), 2594–2605.

[132] A. Kovtun, S. Bergdolt, Y. Hägele, R. Matthes, J. D. Lambris, M. Huber-Lang, and A. Ignatius. Complement receptors C5aR1 and C5aR2 act differentially during the early immune response after bone fracture but are similarly involved in bone repair. *Scientific Reports* 7.1 (2017), 1–12.

[133] A. Ignatius, C. Ehrnthaller, R. E. Brenner, L. Kreja, P. Schoengraf, P. Lisson, R. Blakytny, S. Recknagel, L. Claes, F. Gebhard, J. D. Lambris, and M. Huber-Lang. The anaphylatoxin receptor C5aR is present during fracture healing in rats and mediates osteoblast migration in vitro. *The Journal of Trauma* 71.4 (2011), 952.

[134] S. L. Dallas, Y. Xie, L. A. Shiflett, and Y. Ueki. Mouse Cre models for the study of bone diseases. *Current Osteoporosis Reports* 16.4 (2018), 466–477.

[135] The Jackson Laboratory. Strain 000664. 2023. URL: https://www.jax.org/strain/000664 (visited on 10/16/2023).

[136] S. C. Grubb, G. A. Churchill, and M. A. Bogue. A collaborative database of inbred mouse strain characteristics. *Bioinformatics* 20.16 (2004), 2857–2859.

[137] C. L. Smith, C.-A. W. Goldsmith, and J. T. Eppig. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology* 6.1 (2005), 1–9.

[138] L. Donahue, W. Beamer, M. Bogue, and G. Churchill. Characterization of skeletal geometry and bone strength in 10 inbred strains of mice. 2004. URL: https://phenome.jax.org/projects/Donahue2 (visited on 08/10/2021).

[139] W. Beamer. Bone mineral density and IGF-1 levels in F2 progeny from a C57BL/6J x C3H/HeJ intercross. 2001. URL: https://phenome.jax.org/projects/Beamer2 (visited on 08/10/2021).

[140] J. E. Wergedal, C. L. Ackert-Bicknell, S. W. Tsaih, M. H. Sheng, R. Li, S. Mohan, W. G. Beamer, G. A. Churchill, and D. J. Baylink. Bone Mineral Density (BMD) and related bone phenotypes in a NZB/BINJ and RF/J F2 intercross. 2006. URL: https://phenome.jax.org/projects/Wergedal1 (visited on 08/10/2021).

[141] D. T. Burke, K. M. Kozloff, S. Chen, J. L. West, J. M. Wilkowski, S. A. Goldstein, R. A. Miller, and A. T. Galecki. Dissection of complex adult traits in a mouse synthetic population. *Genome Research* 22.8 (2012), 1549–1557.

[142] The Jackson Laboratory. Nomenclature tutorial. 2023. URL: https://www2.jax.org/nomenclature-tutorial/ (visited on 04/04/2023).

[143] S. R. Goodyear and R. M. Aspden. Mechanical properties of bone ex vivo. *Bone Research Protocols*. Humana Press, 2012, 555–571.

[144] R. Levy, R. F. Mott, F. A. Iraqi, and Y. Gabet. Collaborative Cross mice in a genetic association study reveal new candidate genes for bone microarchitecture. *BMC Genomics* 16.1 (2015), 1–14.

[145] E. J. Chesler, D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson, V. M. Philip, B. H. Voy, C. T. Culiat, D. W. Threadgill, R. W. Williams, G. A. Churchill, D. K. Johnson, and K. F. Manly. The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mammalian Genome* 19.6 (2008), 382–389.

[146] D. J. Schad, B. Nicenboim, P.-C. Bürkner, M. Betancourt, and S. Vasishth. Workflow techniques for the robust use of bayes factors. *Psychological Methods* (2022).

[147] B. Voelkl, H. Würbel, M. Krzywinski, and N. Altman. The standardization fallacy. *Nature Methods* 18.1 (2021), 5–7.

[148] J. Cohen. Statistical power analysis for the behavioral sciences. 2nd ed. New York: Routledge, 1988.

[149] Stan Development Team. Stan user's guide. Version 2.33. 2023. URL: https://mc-stan.org/docs/stan-users-guide (visited on 11/15/2023).

[150] Stan Development Team. RStan: the R interface to Stan. R package version 2.26.23. 2023. URL: https://mc-stan.org/ (visited on 11/15/2023).

[151] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* PAMI-6.6 (1984), 721–741.

[152] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B* 195.2 (1987), 216–222.

[153] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15.1 (2014), 1593–1623.

[154] M. L. Delignette-Muller and C. Dutang. fitdistrplus: an R package for fitting distributions. *Journal of Statistical Software* 64.4 (2015), 1–34.

[155] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke. bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software* 4.40 (2019), 1541.

[156] A. Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software* 16.9 (2006), 1–16.

[157] J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician* 54.3 (2000), 217–224.

[158] Microsoft Corporation. Microsoft Excel. 2016. URL: `https://office.microsoft.com/excel`.

[159] The Mouse Phenotype Database Integration Consortium. Integration of mouse phenome data resources. *Mammalian Genome* 18.3 (2007), 157–163.

[160] K. Wartolowska, A. Judge, S. Hopewell, G. S. Collins, B. J. F. Dean, I. Rombach, D. Brindley, J. Savulescu, D. J. Beard, and A. J. Carr. Use of placebo controls in the evaluation of surgery: systematic review. *British Medical Journal* 348 (2014).

[161] S. S. Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods* 8.26 (2009).

[162] H. M. Vesterinen, E. S. Sena, K. J. Egan, T. C. Hirst, L. Churolov, G. L. Currie, A. Antonic, D. W. Howells, and M. R. Macleod. Meta-analysis of data from animal studies: a practical guide. *Journal of Neuroscience Methods* 221 (2014), 92–102.

[163] I. K. S. Boboc, A. D. Rotaru-Zavaleanu, D. Calina, C. V. Albu, B. Catalin, and A. Turcu-Stiolica. A preclinical systematic review and meta-analysis of behavior testing in mice models of ischemic stroke. *Life* 13.2 (2023).

[164] D. W. Heck. A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology* 72.2 (2019), 316–333.

[165] M. C. Simmonds, J. P. Higginsa, L. A. Stewartb, J. F. Tierneyb, M. J. Clarke, and S. G. Thompson. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials* 2.3 (2005), 209–217.

[166] R. D. Riley, P. C. Lambert, and G. Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal* 340 (2010), c221.

[167] G. B. Stewart, D. G. Altman, L. M. Askie, L. Duley, M. C. Simmonds, and L. A. Stewart. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLoS ONE* 7.10 (2012), e46042.

[168] T. P. Debray, K. G. Moons, G. van Valkenhoef, O. Efthimiou, N. Hummel, R. H. Groenwold, J. B. Reitsma, and G. M. R. Group. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Research Synthesis Methods* 6.4 (2015), 293–309.

[169] A. Tobías, M. Saez, and M. Kogevinas. Meta-analysis of results and individual patient data in epidemiologal studies. *Journal of Modern Applied Statistical Methods* 3.18 (2004).

[170] T. Stijnen, T. H. Hamza, and P. Özdemir. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* 29.29 (2010), 3046–3067.

[171] T. P. Debray, K. G. Moons, G. M. A. Abo-Zaid, H. Koffijberg, and R. D. Riley. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS ONE* 8.4 (2013), e60650.

[172] C. H. Schmid, P. C. Stark, J. A. Berlin, P. Landais, and J. Lau. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology* 57.7 (2004), 683–697.

[173] C. T. Smith, P. R. Williamson, and A. G. Marson. An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes. *Journal of Evaluation in Clinical Practice* 11.5 (2005), 468–478.

[174] M. C. Simmonds and J. P. Higgins. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine* 26.15 (2007), 2982–2999.

[175] S. Witte, H. Schmidli, A. O'Hagan, and A. Racine. Designing a non-inferiority study in kidney transplantation: a case study. *Pharmaceutical Statistics* 10.5 (2011), 427–432.

[176] B. K. Guenhan, T. Friede, and L. Held. A design-by-treatment interaction model for network meta-analysis and meta-regression with integrated nested Laplace approximations. *Research Synthesis Methods* (2018), 179–194.

[177] H. Schmidli, S. Wandel, and B. Neuenschwander. The network meta-analytic-predictive approach to non-inferiority trials. *Statistical Methods in Medical Research* 22.2 (2013), 219–240.

[178] I. K. S. Boboc, A. D. Rotaru-Zavaleanu, D. Calina, C. V. Albu, B. Catalin, and A. Turcu-Stiolica. A preclinical systematic review and meta-analysis of behavior testing in mice models of ischemic stroke. *Life* 13.2 (2023), 567.

[179] S. van der Mierden, C. H. Leenaars, E. C. Boyle, F. L. Ripoli, P. Gass, M. Durst, V. C. Goerlich-Jansson, P. Jirkof, L. M. Keubler, and L. Lewejohann. Measuring endogenous corticosterone in laboratory mice – a mapping review, meta-analysis, and open source database. *ALTEX-Alternatives to Animal Experimentation* 38 (2021), 111–122.

[180] A. A. Besson, M. Lagisz, A. M. Senior, K. L. Hector, and S. Nakagawa. Effect of maternal diet on offspring coping styles in rodents: a systematic review and meta-analysis. *Biological Reviews* 91.4 (2016), 1065–1080.

[181] J. Colomb and Y. Winter. Creating detailed metadata for an R Shiny analysis of rodent behavior sequence data detected along one light-dark cycle. *Frontiers in Neuroscience* 15 (2021), 742652.

[182] S. Weber. Meta-analytic-predictive priors for variances. 2023. URL: https://cran.r-project.org/web/packages/RBesT/vignettes/variances_MAP.html (visited on 10/10/2023).

[183] H. Qi, D. Rizopoulos, and J. van Rosmalen. Sample size calculation for clinical trials analyzed with the meta-analytic-predictive approach. *Research Synthesis Methods* 14.3 (2023), 396–413.

[184] A. J. Sutton, N. J. Cooper, D. R. Jones, P. C. Lambert, J. R. Thompson, and K. R. Abrams. Evidence-based sample size calculations based upon updated meta-analysis. *Statistics in Medicine* 26.12 (2007), 2479–2500.

[185] D. Lakens. Sample size justification. *Collabra: Psychology* 8.1 (2022), 33267.

[186] U. Conradi and A. R. Joffe. Publication bias in animal research presented at the 2008 Society of Critical Care Medicine Conference. *BMC Research Notes* 10.1 (2017), 1–11.

[187] Berlin Institute of Health at Charité. Extent, predictors, and management of publication bias in animal research (EMBARC). 2021. URL: https://www.bihealth.org/de/translation/innovationstreiber/quest-center/projekte/translationale-bioethik/embarc (visited on 11/03/2021).

[188] E. S. Sena, H. B. van Der Worp, P. M. Bath, D. W. Howells, and M. R. Macleod. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS biology* 8.3 (2010), e1000344.

[189] M. Danziger, A. Collazo, U. Dirnagl, and U. Toelch. Balancing sensitivity and specificity in preclinical research. *bioRxiv* (2022).

[190] J. Kruschke. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. 2nd ed. London: Academic Press, 2015.

[191] F. D. Schönbrodt, E.-J. Wagenmakers, M. Zehetleitner, and M. Perugini. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods* 22.2 (2017), 322–339.

[192] D. G. Altman and J. M. Bland. Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal* 311.7003 (1995), 485.

[193] Z. Dienes. Using Bayes to get the most out of non-significant results. *Frontiers in Psychology* 5 (2014).

[194] M. Linde, J. N. Tendeiro, R. Selker, E.-J. Wagenmakers, and D. van Ravenzwaaij. Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the Bayes factor. *Psychological Methods* 28.3 (2021), 740–755.

[195] P. Schwaferts and T. Augustin. Bayesian Decisions using Regions of Practical Equivalence (ROPE): Foundations. 2020. URL: https://epub.ub.uni-muenchen.de/74222/ (visited on 10/10/2022).

[196] S. Gsteiger, B. Neuenschwander, F. Mercier, and H. Schmidli. Using historical control information for the design and analysis of clinical trials with overdispersed count data. *Statistics in Medicine* 32.21 (2013), 3609–3622.

[197] L. Di Scala, J. Kerman, and B. Neuenschwander. Collection, synthesis, and interpretation of evidence: a proof-of-concept study in COPD. *Statistics in Medicine* 32.10 (2013), 1621–1634.

[198] A. Biswas and J.-F. Angers. Discussion on "Predictively consistent prior effective sample sizes", by Beat Neuenschwander, Sebastian Weber, Heinz Schmidli, and Anthony O'Hagan. *Biometrics* 76.2 (2020), 591–594.

[199] C. G. Begley and L. M. Ellis. Raise standards for preclinical cancer research. *Nature* 483.7391 (2012), 531–533.

[200] E. Loken and A. Gelman. Measurement error and the replication crisis. *Science* 355.6325 (2017), 584–585.

[201] S. N. Goodman, D. Fanelli, and J. P. Ioannidis. What does research reproducibility mean? *Science Translational Medicine* 8.341 (2016).

[202] W. H. Kruskal. Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association* 52.279 (1957), 356–360.

[203] J. van Doorn, A. Ly, M. Marsman, and E.-J. Wagenmakers. Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics* 47.16 (2020), 2984–3006.

[204] L. Leon-Novelo, B. N. Bekele, P. Müller, F. Quintana, and K. Wathen. Borrowing strength with nonexchangeable priors over subpopulations. *Biometrics* 68.2 (2012), 550–558.

[205] P. Müller and R. Mitra. Bayesian nonparametric inference – why and how. *Bayesian analysis* 8.2 (2013), 269–302.

[206] P.-C. Bürkner and M. Vuorre. Ordinal regression models in psychology: a tutorial. *Advances in Methods and Practices in Psychological Science* 2.1 (2019), 77–101.

[207] F. Konietschke, K. Schwab, and M. Pauly. Small sample sizes: A big data problem in high-dimensional data analysis. *Statistical Methods in Medical Research* 30.3 (2021), 687–701.

[208] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5.2 (2012), 189–211.

[209] E. Brunner, A. C. Bathke, and F. Konietschke. Rank and pseudo-rank procedures for independent observations in factorial designs. Cham: Springer, 2018.

[210] D. J. Nott, X. Wang, M. Evans, and B.-G. Englert. Checking for prior-data conflict using prior-to-posterior divergences. *Statistical Science* 35.2 (2020), 234–253.

[211] K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, S. Micallef, S. Roychoudhury, and L. Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics* 13.1 (2014), 41–54.

[212] D. J. Nott, M. Seah, L. Al-Labadi, M. Evans, H. K. Ng, and B.-G. Englert. Using prior expansions for prior-data conflict checking. *Bayesian Analysis* 16.1 (2021), 203–231.

[213] M. M. Lovric. Conflicts in Bayesian statistics between inference based on credible intervals and Bayes factors. *Journal of Modern Applied Statistical Methods* 18.1 (2020), eP3320.

[214] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* 21 (2014), 1157–1164.

[215] R. Kelter. The evidence interval and the Bayesian evidence value: on a unified theory for Bayesian hypothesis testing and interval estimation. *British Journal of Mathematical and Statistical Psychology* 75.3 (2022), 550–592.

[216] J. M. Hancock, P. N. Schofield, C. Chandras, M. Zouberakis, V. Aidinis, D. Smedley, N. Rosenthal, and K. Schughart. CASIMIR: Coordination and Sustainability of International Mouse Informatics Resources. *8th IEEE International Conference on BioInformatics and BioEngineering*. 2008, 1–6.

[217] Fraunhofer ITEM Hannover. RITA Registry of Industrial Toxicology Animal-data. 2022. URL: https://reni.item.fraunhofer.de/reni/public/rita (visited on 05/04/2021).

[218] F. Sanz, F. Pognan, T. Steger-Hartmann, C. Díaz, M. Cases, M. Pastor, P. Marc, J. Wichard, K. Briggs, T. Watson David Kleinöder, C. Yang, A. Amberg, M. Beaumont, A. J. Brookes, S. Brunak, M. T. D. Cronin, G. F. Ecker, S. Escher, N. Greene, A. Guzmán, A. Hersey, P. Jacques, L. Lammens, J. Mestres, W. Muster, H. Northeved, M. Pinches, J. Saiz, N. Sajot, A. Valencia, J. van der Lei, N. P. E. Vermeulen, E. Vock, G. Wolber, and I. Zamora. Legacy data sharing to improve drug safety assessment: the eTOX project. *Nature Reviews Drug Discovery* 16.12 (2017), 811–812.

[219] F. Pognan, T. Steger-Hartmann, C. Díaz, N. Blomberg, F. Bringezu, K. Briggs, G. Callegaro, S. Capella-Gutierrez, E. Centeno, J. Corvi, P. Drew, W. C. Drewe, J. M. Fernández, L. I. Furlong, E. Guney, J. A. Kors, M. A. Mayer, M. Pastor, J. Piñero, J. M. Ramírez-Anguita, F. Ronzano, P. Rowell, J. Saüch-Pitarch, A. Valencia, B. van de Water, J. van der Lei, E. van Mulligen, and F. Sanz. The etransafe project on translational safety assessment through integrative knowledge management: achievements and perspectives. *Pharmaceuticals* 14.3 (2021), 237.

[220] F. Schönbrodt. BFDA: Bayes factor design analysis package for R. 2016. URL: https://github.com/nicebread/BFDA (visited on 08/04/2023).

Danksagung

Ich möchte mich herzlich bei meinem Betreuer Prof. Dr. Benjamin Mayer und bei Prof. Dr. Rainer Muche für die Ermöglichung und Unterstützung meiner Arbeit bedanken. Weiterhin möchte ich mich bei meinen Kollegen für die gute Zusammenarbeit und das angenehme Arbeitsumfeld bedanken. Außerdem möchte ich mich bei Prof. Dr. Melanie Haffner-Luntzer vom Institut für Unfallchirurgische Forschung und Biomechanik der Universität Ulm für die Bereitstellung der Tierdaten und für die Diskussion über den Nutzen von historischen Kontrolldaten bedanken. Weiterhin möchte ich mich für die Unterstützung des Staates Baden-Württemberg durch das bwHPC Rechencluster bedanken.

Schließlich möchte ich auch meinen Eltern und Freunden für ihre Unterstützung danken.

Curriculum Vitae

Der Lebenslauf wurde aus Gründen des Datenschutzes entfernt.

Der Lebenslauf wurde aus Gründen des Datenschutzes entfernt.

Publikationsliste

Unseld, T. Designing translational animal experiments by Bayesian MAP approaches. 2023. URL: https://arxiv.org/abs/2301.05572.

Unseld T., Ruckerbauer, L., and Mayer, B. Permutation tests are able to outperform common statistical hypothesis tests in small samples. 2023. *(Manuskript eingereicht)*

Unseld, T., Bluhmki, T., Beyersmann, J., Beck, E., Padberg, S., Stegherr, R. Landmarking for left-truncated competing risks data. 2023. *(Manuskript eingereicht)*