arXiv:2301.05572v1 [stat.ME] 13 Jan 2023

通过贝叶斯荟萃分析预测方法设计转化动物实验

预印本

Theresa Unseld*

德国乌尔姆乌尔姆大学流行病学与医学统计学系

2023年1月16日

摘要

在欧盟，动物实验的规划和实施受到严格的法律约束。尽管如此，许多临床前动物实验的设计仍不够完善。因此，在一项动物实验中取得的发现，可能无法在另一项动物实验中重现；转化动物研究中的发现也可能无法转化至人类。在经典频率学派框架下设计新实验时，针对零假设的统计检验、显著性阈值和最低相关效应量给定后，研究者会选择新实验的样本量，以期至少达到一定的统计功效。统计检验是数据的函数，用于根据数据对其背后未观测到的目标参数进行统计推断。在贝叶斯框架中，推断结合了新观测数据提供的信息，以及代表参数先验信息的先验分布。在转化动物实验中，先验信息来自此前在相似动物中针对相同结局开展的实验。通过将历史数据汇总为（贝叶斯）荟萃分析模型，并利用该模型预测新实验的数据，可以系统地将先验信息纳入新动物实验的设计与分析。这种方法称为荟萃分析预测（MAP）方法。本文介绍如何运用 MAP 方法设计转化动物实验的相关概念，并将其与经典频率学派功效导向的样本量规划进行比较。本文还探讨了这些方法在转化动物研究实际应用中的机遇与挑战，并特别关注先验分布的构建以及通过设计分析进行样本量计算。相关讨论以一个真实世界的转化研究案例为引。

关键词 转化 化研究、贝叶斯统计学、荟萃分析、设计分析、样本量计算 n

1 引言

转化研究是推动人类医学中新方法和疗法发展的关键环节。转化研究处于临床前研究的后期阶段，其目标是将基础实验室临床前研究的发现转化为临床应用，用于治疗人类疾病。在转化动物研究中，研究人员会考察与临床相关表型或病理状况有关的生物学通路。这些通路通常是复杂的构成，其机制无法通过单项研究实验充分揭示。因此，转化研究的成功从根本上取决于能否审慎地结合既往见解和积累的专业知识，思考实验带来的新认识。在贝叶斯分析中，先验知识以概率分布的形式正式纳入对新观测数据的分析，并更新为后验分布，随后用于

*theresa.unseld@uni-乌尔姆.de

贝叶斯推断。已有多位作者强调了贝叶斯统计学在临床前研究中的价值（例如参见 Spiegelhalter 等人（2004年）、Walley 等人（2016年）、Kramer 与 Font（2017年）、Bonapersona 等人（2019年）、Yang 与 Novick（2019年）、Gelman 与 Vákár（2019）以及 Novick 与 Zhang（2021年））。尽管如此，贝叶斯方法在临床前动物实验的规划和分析中的实际应用仍很少见（参见 Walley 等人（2016年）、Kramer 与 Font（2017年）以及 Bonapersona 等人（2019年））。应用贝叶斯方法的一项挑战是如何设定先验分布。转化动物实验通常具有这样的特点：出于动物福利考虑，实验各组的样本量都较小（参见 Mayer 等人（2018年）），这与 Russel & Burch（1959年）提出的 3R原则相符。尤其在这种情况下，先验分布的选择会对后验推断产生重大影响。设定能够准确反映先验信息的良好先验分布，有助于稳定后验推断、得出更精确的估计值、降低单个极端观测值的影响，并指出新数据的测量值中是否可能存在导致后验分布较宽的异常或意外情况。然而，如果没有透明的依据来说明先验的选择，也未充分理解先验分布对后验推断的影响，贝叶斯分析的最终结论就可能无法（充分）敏锐地反映新观测数据中的证据。另一方面，如果选择的先验分布对后验推断的影响几乎为零，相较于新实验中的数据而言，那么贝叶斯推断的一项重要特征及其相较于频率学派推断的相关优势便不复存在。

一种系统设定先验分布的方法，是根据相关文献或数据库的荟萃分析来指定先验分布，Neuenschwander 等人（2010）、Pullenayegum（2011）、Rhodes 等人（2015）以及 Bartoš 等人（2021）均提出了这一方法。荟萃分析是一种常用工具，用于在统一的统计模型中汇总多个统计实验的信息，并量化不同实验之间的变异性。作为一项要求，所有实验都必须针对同一个问题。这一假设对于来自一系列针对相同表型结局的实验的对照组而言通常是合理的。荟萃分析模型可用于估计新实验的预测分布。这些预测分布可用于为新实验的分析推导先验分布。这种先验设定以及新实验的贝叶斯规划和分析方法被称为荟萃分析预测方法（MAP），见 Neuenschwander 等人（2010）。这些 MAP先验属于基于数据的先验，正是本文的重点。Mikkola 等人（2021）总结了其他先验设定方法，本文将在讨论部分对其进行简要讨论。Neuenschwander 等人（2010）以临床研究对照组的均值为例阐释了 MAP 方法。然而，他们并未利用历史数据为方差等其他模型参数推导先验分布。此外，动物实验具有自身的挑战，人体临床试验中提出的方法在应用于动物实验时，若不考虑可能需要的调整，不能直接照搬（见 Walley 等人（2016）、Kramer & Font（2017））。首先，由于实验各组的样本量较小，动物实验所得的估计值通常具有较大的不确定性。另一项挑战在于，尽管开展了许多动物实验，但分析结果通常缺乏整理，且访问受到限制（见 Kramer & Font（2017）、Bonapersona 等人（2019）、Novick & Zhang（2021））。人们一直设想建立检索工具，以对临床前研究开展系统综述（见 Bahor 等人（2021）），或建立共享的大型数据库，以汇集更多机构的信息（见 Keenan 等人（2009）、Beckers 等人（2009）、Maddatu 等人（2012）、McEntyre 等人（2015）、Steger-Hartmann 等人（2020）、Pognan 等人（2021））。在这些数据库完全建成之前，剩余的可行选择是利用现有的少量历史信息构建先验分布，并意识到模型参数真实值存在当前不确定性。贝叶斯方法能够在仅有少量小型既往实验这一具有挑战性的情形下拟合荟萃分析模型。Friede 等人（2017*a*）最近提出了在少量小型研究背景下将贝叶斯荟萃分析应用于人类研究的方法。此外，Walley 等人（2016）还讨论了在动物实验背景下应用 MAP 方法。然而，目前关于如何使用此类贝叶斯荟萃分析预测方法分析临床前转化研究中的真实案例，以及如何在贝叶斯框架下开展动物实验样本量计算的实例仍然很少（见 Kramer & Font（2017））。

在欧盟，动物实验的规划和实施受到严格法律条件的约束，例如《欧盟指令 2010/63/EU》对此作出了规定；在德国，相关规定见动物福利立法《动物保护实验动物条例》（参见 Bundesministerium der Justitz（2010））。尤其是，动物实验必须获得相应主管机关的批准。研究人员可通过提交动物实验申请书来申请批准。如果动物实验旨在检验假设，则申请书必须附上表格或生物统计学审查报告，其中应说明分析方法，并论证所提议的样本量。传统上，样本量计算采用频率学派框架：给定一个包含其参数的数据模型，选择能够使针对零假设的统计检验达到预设检验效能的最少动物数量。然而，针对零假设的经典显著性检验（零假设显著性检验，NHST）和功效分析方法，因多种原因受到批评（参见 Gelman & Carlin（2014）、Kruschke（2015）、Schönbrodt & Wagenmakers（2018）、Stefan 等人（2019）、Gelman、Hill 与 Vehtari（2020））。通过频率学派检验某一假设并宣称其具有统计显著性，等同于根据相应检验的

p值或置信区间宣称统计显著性（参见 Lehmann & Romano（2006））。Wagenmakers 等人（2020）、Gelman、Hill 与 Vehtari（2020）以及 Schad 等（2022）等人批评了这些用于检验零假设的决策规则。其中一项批评与预测无关性原则 有关（参见 Wagenmakers 等人（2020））；贝叶斯因子也被提出作为替代方案。Gelman & Carlin（2014）以及 Gelman、Hill 与 Vehtari（2020）指出，为了达到某个统计检验的特定检验效能而选择实验设计（同时限制检验发生第一类错误的概率），会忽略实验设计的其他重要目标。具体而言，他们认为，若把重点放在检验效能和统计显著性上，报告的估计值就会出现系统性偏倚。这是因为，实际上较小的效应只有在观测到的效应量碰巧（因偶然因素）较大时才会被报告，而这类情况极易导致效应被高估，甚至估计出错误的方向。经典框架下的样本量计算基于效应量和其他模型参数的固定估计值。一种做法是根据研究人员此前的实验推导这些估计值。但这种做法存在问题，因为动物实验中的样本量较小，导致参数估计值伴有很大的不确定性（参见 Mayer & Muche（2013）），而单个点估计无法充分反映这些不确定性。因此，动物实验中的样本量计算往往作用有限。作为以检验效能为导向的样本量计算的替代方案，Kruschke（2015）以及 Gelman、Hill 与 Vehtari（2020）提出了采用伪数据模拟进行设计分析。本文在转化性动物实验样本量计算的背景下，综合考量贝叶斯分析、荟萃分析、采用预测方法进行先验设定，以及通过伪数据模拟进行设计分析。据作者所知，目前尚无研究综合运用这些方法来说明转化性动物实验的样本量计算。继本节引言之后，第二节介绍经典频率学派框架下的样本量计算，并介绍作为替代方案的伪数据设计分析。此外，还阐述了如何运用基于历史数据的贝叶斯最大后验方法来设计和分析新实验。本文采用分布型贝叶斯模型，以处理历史数据和新数据中组间方差不相等的问题。第三节将上述考量应用于一个实际案例。第四节总结主要观点，并讨论可能的扩展。

2 方法

2.1 统计假设 检验与效能导向的样本量计算 布

在进行样本量计算时，必须对所有实验组的分布模型以及实验设计作出假设，后者需明确要进行哪些比较。此外，还必须定义感兴趣的可估计效应 δ，通常代表两个或多个组之间的均值差异或效应差异。此外，为了作出二元决策，判断数据中是否有充分证据表明感兴趣的真实效应 δ 不等于零效应 δ_0，需要一个决策函数。在经典频率学派框架中，决策函数是统计假设检验 φ。统计假设检验 φ 将检验统计量 T 与临界值 c 进行比较，以决定是否拒绝零假设

$$H_0 : \{\delta = \delta_0\} \text{vs. } H_1 : \{\delta \neq \delta_0\} \tag{1}$$

如果 H_0 被拒绝，则称备择假设 H_1 被接受。检验统计量 T 是根据数据构造的函数，其在 H_0 下的概率分布已知，或可通过已知分布近似得到。一种常见情形是比较两个独立组，即实验组（E）和对照组（C）（例如基因敲除动物与野生型动物），并假设数据服从正态分布：

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

2.1.1 频率学派分析中的假设检验

作为针对均值差的频率学派假设检验，在方差相等 $\sigma_E = \sigma_C$（同方差性）的情况下，可以使用双样本 t 检验 $\delta = |\theta_E - \theta_C|$（参见 Lehmann & Romano (2022)）。然而，由于处理效应存在变化，实验组的残差方差通常大于对照组，因此 Welch (1938)、Satterthwaite (1941) 提出的 Welch 检验（Satterthwaite 检验）更为合适，因为它不要求满足同方差性假设：

$$T_{\text{Welch}} = \frac{\overline{y}_1 \cdot - \overline{y}_2 \cdot}{\sqrt{\frac{\sigma_C^2}{n_1} + \frac{\sigma_E^2}{n_2}}} \tag{3}$$

在 H_0 下，检验统计量 T_{Welch} 的分布可用自由度经过修正的 t 分布 $t_{\tilde{\nu}}$ 近似 $\tilde{\nu}$（详情参见 Welch (1938, 1947)）。方差 σ_E^2 和 σ_C^2 可根据观测数据估算为经验方差。相应的双侧假设检验为

$$\varphi_{\text{Welch}} = 1\{|T_{\text{Welch}}| > t_{\tilde{\nu};1-\frac{\alpha}{2}}\} \tag{4}$$

其中，$t_{\tilde{\nu};1=\frac{\alpha}{2}}$ 表示自由度经过修正的 t分布的 $1-\frac{\alpha}{2}$ 分位数。假设检验可能产生两类标准错误：

1. 第一类错误：φ 拒绝 H_0，尽管它为真。2. 第二类
错误：φ 不拒绝 H_0，尽管它不为真

这两类错误无法同时降低。由于通常认为第一类错误比第二类错误更严重，因此经典假设检验 $\varphi = \varphi_\alpha$ 会设定显著性水平 α，以限制第一类错误的概率；随后，在所有显著性水平为 α 的检验中，通过最小化第二类错误的概率来寻找最优检验（参见 Lehmann & Romano（2006年））。由此得到检验 φ_α^*，它在将第一类错误控制在水平α的同时，具有最大的检验效能 P（$\varphi = 1|H_1$）。随后，经典方法会将样本量计算为能够使所选假设检验 φ_α^* 以至少 $1-\beta$的概率检测到具有临床意义的效应量 δ^* 的最小样本数。更具体地说，在检验统计量具有简单解析分布形式的应用中，可以通过求解检验效能不等式来计算样本量

$$P(\varphi_{\text{Welch},\alpha,n_E,n_C} = 1|\delta \geq \delta^*) \geq 1-\beta. \tag{5}$$

对于 Welsh 检验，临界值 $t_{\tilde{\nu};1-\frac{\alpha}{2}}$ 本身是样本量的函数，这是通过（近似的）自由度体现的。因此，无法直接求解该不等式。可采用的替代方法包括用正态分布近似临界值，或使用迭代方法。

2.1.2 贝叶斯分析中的假设检验

在频率学派语境中，“统计显著”意味着检验统计量 T大于临界值 c（此处为：$c = t_{\nu;1-\frac{\alpha}{2}}$），或者等价地说，检验的 p 值小于 α，或效应 δ在 $1-\alpha$ 水平上的置信区间不包含零值 δ_0（见 Lehmann & Romano（2006年））。在贝叶斯框架下，也可以建立类似的决策规则或“检验”。在完整的贝叶斯模型中，所有参数都被建模为概率分布，并各自具有先验分布。贝叶斯分析的理念是利用新数据中的信息，通过贝叶斯定理将这些先验分布更新为后验分布。关于先验分布的详情见第 2.4 节。为所有模型参数设定概率分布，而不是将参数固定为具体数值，一方面能够比频率学派框架更好地表征不确定性；另一方面，这通常会导致复杂的后验密度函数，需要计算高维积分，且无法再用解析形式表示。使用马尔可夫链蒙特卡罗方法进行数值计算是一种常见的替代方法。在马尔可夫链蒙特卡罗方法中，参数分布的后验分布由一系列 MCMC 抽样值近似得到。对于单侧检验问题，例如上侧检验问题 H_0： $\{\delta \leq \delta_0\}$ 对比 H_1： $\{\delta > \delta_0\}$，可以计算在给定数据 $y = (y_C, y_E)$的条件下，δ 大于零值 δ_0或临床相关值 δ^* 的后验概率 $p(\delta|y)$。为了将这一理念转化为决策规则，可以将该后验概率与预先设定的临界值进行比较，例如 Weber等人（2021年）所建议的临界值。当被检验效应 δ 由连续概率分布建模时，这种方法无法用于双侧问题（1）；例如，$\delta = |\theta_E - \theta_C|$ 表示两个连续均值之差时便是如此。在这种情况下，单点事件 $P(\{\delta = \delta_0\})$的后验概率等于零。替代方法包括考察双侧贝叶斯因子或贝叶斯可信区间。

可信区间是频率学派置信区间的贝叶斯版本，但两者的解释略有不同。根据定义，针对 δ 的频率学派 $1-\alpha$ 置信区间是一个具有随机边界的区间，它以 $1-\alpha$ 的概率覆盖未知（但视为固定）的参数 δ。因此，如果根据相应的概率模型生成数据，并为每个数据集计算一个置信区间，那么预期这些区间中有 $(1-\alpha)$% 会覆盖（固定的） δ。而 $1-\alpha$ 贝叶斯可信区间的解释是：将其构造为边界固定的区间，使 δ 的一次随机实现落在可信区间内的概率为 $1-\alpha$。等价地说，给定数据 y，贝叶斯 $1-\alpha$ 可信区间包含 δ 的后验概率质量的 $(1-\alpha)$%（见 Held & Sabanés Bové（2014））。然而，这一定义并不能唯一确定区间。贝叶斯可信区间有两种常见类型：分位数区间和最高密度区间。其边界

参数的 $1-\alpha$ 分位数区间的边界由其后验分布的 $\frac{\alpha}{2}$ 和 $1-\frac{\alpha}{2}$ 分位数给出，也称为等尾区间。在实际应用中，分位数区间的边界可以用马尔可夫链蒙特卡洛抽样值的分位数近似。最高密度区间（HDI）定义为这样一种可信区间：区间内每个点的后验密度都高于区间外任意一点的后验密度。这是作为分布摘要的一项理想特性，在偏态分布中尤其重要，因为对于分位数区间，区间内的参数值可能比区间外的参数值概率更低。此外，HDI 是所有可能的可信区间中最短的；当用于后验推断时，这是一项理想特性。另一方面，分位数区间的优势在于，对于变换参数而言，它比 HDI 更容易解释。这是因为，只需对为变换参数推导出的区间进行反变换，就能得到变换参数的分位数区间。相比之下，无法通过对变换参数的 HDI 进行简单反变换，来推导未变换参数的 HDI。另一个区别是，分位数区间总是包含后验分布的中位数，而 HDI 总是包含后验分布的众数（Kruschke（2015））。在对称分布中，分位数区间和 HDI 得出的结果相似。此外，在某些条件下，贝叶斯可信区间也与频率学派置信区间一致（见 Jaynes & Kempthorne（1976））。使用贝叶斯可信区间，可以通过检验 δ_0 是否落在 δ 的 $1-\alpha$ 后验可信区间内来设定决策规则。一种可能更有意义的方法是检验后验区间是否排除了实际等效区间（ROPE）；该区间可定义为所有绝对值小于最小临床相关效应量 δ_{rel} 的参数值（Kruschke（2015）建议如此）。作为另一种替代方案，Kruschke 建议将目标设为达到一定精度，而非达到某一特定检验效能。相应的决策规则可以通过判断可信区间的宽度是否小于代表目标精度的阈值来实施。

零假设的经典决策规则因零假设检验而受到批评，例如 Wagenmakers 等人（2020）、Gelman、Hill 和 Vehtari（2020）以及 Schad 等人（2022）的批评。其中一项批评与预测无关性原则有关。该原则指出，如果数据在零模型 $\mathcal{M}_0$（对应H_0）和备择模型 $\mathcal{M}_1$（对应 H_1）下得到同等程度的预测，那么这些数据——Wagenmakers 等人（2020）将其称为无信息或无关数据——就不应导致偏向其中一个模型。然而，在上述零假设显著性检验（NHST）的情形中，如果区间或 p 值仅在两种模型之一之下进行估计，这种偏向就可能发生。因此，为量化支持异于零假设的效应的证据，人们提出使用贝叶斯因子。贝叶斯因子比较观测数据在（至少）两个模型下的概率 $\mathcal{M}_0$、$\mathcal{M}_1$：

$$\mathbf{BF}_{10} = \frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_0)} \tag{6}$$

这个贝叶斯因子表明，在 H_1 下的模型比在 H_0 下的模型更有可能生成这些数据。它与后验模型概率的赔率 $\frac{p(\mathcal{M}_1|y)}{p(\mathcal{M}_0|y)}$ 相关，因为它是先验模型赔率之比 $\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}$ 在观察数据后发生变化的倍数：

$$\frac{p(\mathcal{M}_1|y)}{p(\mathcal{M}_0|y)} = \mathbf{BF}_{10} \cdot \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}$$

如果贝叶斯因子大于 1，这表明在观测到数据后，与先验预期相比，模型 H_1 相对于模型 H_0 的赔率有所增加。Schad 等（2022）指出，若要使这种方法能够合理地反映数据中观测到的证据，为模型参数选择适当的先验分布至关重要。基于先验分布和数据估计贝叶斯因子的典型方法包括桥接抽样（Bennett（1976年））和 Savage-Dickey 方法（Dickey 与 Lientz（1970））。要制定基于贝叶斯因子的决策规则，必须设定一个阈值，以确定何时拒绝零假设。Lee与Wagenmakers（2014）根据 Jeffreys（1961）调整后，提供了一套贝叶斯因子的粗略解释方案。他们规定，如果贝叶斯因子 BF_{10} 大于 3，则可视为支持 H_1 的中等证据。可以将满足至少中等证据条件的 H_1 的比例计算为二元决策函数。这种分类和决策规则便于概览，也提供了一种根据贝叶斯因子估算类似功效的方法，但不应将其用作严格规则（见 Schad 等（2022）、Schönbrodt 和 Wagenmakers（2018））。相反，应考虑贝叶斯因子的完整分布；理想情况下，还应按照 Schad 等（2022）的介绍，进行基于模拟的校准和敏感性分析，以确保正确解释贝叶斯因子。或者，也可以采用启发式决策规则：如果模型 H_1 的后验模型概率 $p(\mathcal{M}_i|y)$ 较高，则判定支持该模型，$i = 0,\ 1$（见 Schad 等（2022））。然而，这一决策规则未包含对任一假设都缺乏证据的情况；只有当相应决策规则中的两种错误同样严重时，它才是最优的：即当数据实际上对应模型 $\mathcal{M}_0$ 时却判定为 H_1，以及当数据实际上对应

模型 $\mathcal{M}_1$ 时却判定为 。在临床研究或临床前研究中，通常并非如此。推导决策规则的一种更具原则性的方法是定义并优化效用函数。效用函数规定在零假设或备择假设为真的条件下，各种决策的成本或价值；要评估决策函数的表现，就必须使用效用函数。在频率学派框架中，效用以第一类错误率和第二类错误率来定义；决策函数则通过限制第一类错误并针对第二类错误进行优化来构建。以贝叶斯因子为基础的决策规则与频率学派假设检验之间的区别之一是，贝叶斯因子能够区分无证据与支持 H_0 的证据，而频率学派检验则不能。

2.2 基于模拟的设计分析

Gelman & Carlin（2014）、Gelman 与 Vákár（2019）、Gelman、Hill 与 Vehtari（2020）提出使用伪数据模拟进行设计分析，以此作为经典的、以检验效能为导向的样本量计算方法的替代方案。利用伪数据模拟，可以结合相关的候选样本量，考察参数变化对预先设定的决策函数的影响。为了确定是否达到预先设定的统计目标，需要对数据拟合模型、进行统计分析，并评估先前定义的决策函数。统计目标通过上一节介绍的效用函数形式化。在此，我们通过计算以下情况发生的百分比，将效用表示为I类和II类错误率，或假发现率（FDR）和真发现率（TDR）：数据模拟时，δ 等于零效应，但决策函数判定为反对零效应；以及数据模拟时，δ 对应不断增大的效应量，而决策函数判定为反对零效应。目标是（找到一个样本量），使真发现率达到特定水平，或检验效能达到特定水平，同时将假发现率或I类错误率控制在某个 α以内。还会考察其他模型特征，例如*S*型（符号）错误，即在结果具有统计显著性的前提下，效应估计值的符号与真实效应的符号相反的概率；以及*M*型（幅度）错误，即在结果具有统计显著性的前提下，效应估计值的绝对值大于真实效应绝对值的概率（Gelman & Carlin，2014）。此外，Gelman 与 Vákár（2019）还考察了在真实效应量不同的情况下，模拟数据中真实效应量估计值的均方误差（MSE）。研究还会将贝叶斯因子的分布可视化，并考察贝叶斯因子错误地大于或小于1的频率、模型在 H_1 下的后验模型概率之95%区间下限是否超过50%，以及贝叶斯因子落在Jeffreys（1961）定义的各类别中的比例。根据效用函数和决策规则，可以选择一个能够达到一个或多个预先设定目标的最小样本量。

$R = 10000$模拟数据集按照模型 (2) 构建，由随机对照组、处理效应和组特异性残差相加得到。在本研究中，新实验的数据是在频率学派框架下生成的，其中 δ、θ_C、ψ 和 λ 取固定值，模拟数据中的随机性仅来自组特异性残差。

$$
\begin{aligned}
&y^*_{ik,r} = \theta^*_{C,r} + \delta^*_r 1(i = E) + \epsilon^*_{il,r} \\
&\epsilon^*_{ik,r} \sim \mathcal{N}(0, \sigma^{*2}_i),\ \sigma^*_i = \exp(\eta^*_{\sigma_{i,r}}),\ \text{where } \eta^*_{\sigma_{i,r}} = \psi^*_r + \lambda^*_r 1(i = E)
\end{aligned}
\tag{7}
$$

对于数据集 $r = 1$，.....，$\tilde{n}_i$ 中组 $i =$C、E 的动物，.....10000。或者，也可以按照讨论部分所述，在完全贝叶斯框架下生成新数据。

对于每个模拟数据集，均使用 R Core Team（2021）中的 brms 软件包的 brms 函数拟合贝叶斯模型（Bürkner，2021年），并采用默认的 MCMC 参数。总体效应的频率学派点估计和置信区间则使用基础 R 中的 lm 函数进行估计。Gelman、Vehtari、Simpson、Margossian、Carpenter、Yao、Kennedy、Gabry、Bürkner 和 Modrák（2020年）以及 Schad 等人（2021年）将贝叶斯框架中通常采取的步骤和决策描述为贝叶斯工作流。拟合贝叶斯模型后，贝叶斯工作流的下一步是验证计算结果。为判断 MCMC 抽样值是否可能已收敛至目标后验分布，需要在收敛诊断中检查 MCMC 链的特征。具体而言，需要检查 MCMC 轨迹图、自相关图、有效样本量和 R 统计量（详情见 Gelman 等（2013））。使用 bayestestR 软件包将贝叶斯可信区间估计为最高密度区间，并将分位数区间计算为后验 MCMC 抽样值的经验分位数。使用 brms 软件包的假设函数计算贝叶斯因子。为处理异方差性，在 brms 中拟合分布模型；在频率学派模型中，则通过使用 sandwich 软件包估计稳健置信区间来处理异方差性。该软件包（Zeileis，2006年）可估计异方差一致（HC）方差。更具体地说，频率学派设计中采用 HC3 型估计量，该估计量也适用于较小样本量（见 Long 和 Ervin，2000年）。此外，还计算 Welsh 检验中的频率学派 p 值。

2.3 历史数据的荟萃分析模型

抽样模型（7）中的分布，以及对模拟数据（和新数据）进行贝叶斯分析时所用的先验分布，均基于相关历史数据的贝叶斯荟萃分析模型。我们希望，通过采用贝叶斯估计和历史证据，能够更好地反映先验知识和不确定性，并找到一个样本量，使其比采用经典频率学派方法更有可能真正实现预先设定的统计目标。在最理想的情况下，如果先验分布正确反映了模拟数据的（主要方面），那么在贝叶斯分析新数据时将历史信息用作先验分布，甚至可以减少达到预先设定的统计目标所需的动物数量。

2.3.1 正态-正态层级模型

历史数据被建模为来自 G 个不同动物实验的数据，每个实验包含 n_g 只动物， $g = 1$、.....、 G。最简单且最常采用的情形是假设数据服从正态分布：

$$y_{gk} = \theta_g + \epsilon_{gk}, \qquad \epsilon_{gk} \sim \mathcal{N}(0, \sigma_g^2) \;\; k = 1, \ldots, n_g, \; g = 1, \ldots, G. \tag{8}$$

其中残差为 ϵ_{gk}、 $g = 1$、.....、 G、 $k = 1$、.....、 n_g，各实验特定的均值为 θ_g、 $g = 1$、.....、 G，并以算术平均数 $\bar{y_g} = \frac{1}{n_g}\sum_{k=1}^{g} y_{gk}$、 $g = 1$、.....、 G 估计。对于不适合采用正态假设的数据，例如偏态数据和非连续数据，通常可以通过变换使其近似服从正态分布，从而使用该模型进行处理（参见 Hedges & Olkin（1985）、Hartung & Knapp（2001）、Higgins & Green（2011））。为使历史数据能够用于新实验的分析，我们假设新数据与历史数据是可交换的，即假定新数据与历史数据之间不存在系统性差异。我们通过对历史数据进行随机效应荟萃分析来建模这一假设，并按照 Neuenschwander 等人（2010）所述，使用该模型均值的先验预测分布为新实验中的参数构建先验分布。历史实验的方差可能因年龄、动物品系、实验室条件或测量仪器不同而产生异质性。这种异质性通过异质性分量 γ_g 建模，它们表示各实验特定均值 μ 、 $\theta_g = \mu + \gamma_g$、 $g = 1$、 ..... 、 G 相对于共同均值 μ 的偏离。感兴趣的实验特定参数 θ_g 的分布中的参数 μ 和 τ 称为超参数。

历史数据的正态分布-正态分布层级模型（NNHM）表示如下

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

该模型称为层级模型，因为它包含两个层级上相互关联的抽样分布；又称正态分布-正态分布层级模型，因为假定个体数据层级的残差 ϵ_{gk} 以及实验特定均值层级的异质性分量 γ_g 均服从正态分布。在频率学派框架中，该模型的正态分布均值 $\theta_g \sim \mathcal{N}(\mu, \tau^2)$也称为随机效应模型。Friede 等人（2017*a*,*b*）证明，与频率学派框架相比，在贝叶斯框架下进行荟萃分析尤其有益，特别是在研究数量少且样本量较小的情况下。层级模型中，以共同均值作为合并效应来估计组特定均值，会得到所谓的收缩估计值；与固定效应模型中对独立均值的估计相比，这些估计值会向共同均值 μ 收缩。收缩估计的优点是，单个极端值的影响会减弱，并且可以通过借用其他组的信息来降低单个组的不确定性。某一组的收缩程度 $g \in \{1, \ldots, G\}$ 取决于其样本量 n_g（或相应的标准误 s_g）以及组间异质性 τ^2。样本量越小 n_g（越大s_g），且异质性越小 τ^2，收缩程度越高（详情参见 Gelman 等（2013）、Schmidli 等人（2014）、Neuenschwander 等人（2016）、Wandel 等人（2017）、Röver & Friede（2021））。

2.4 先验分布

在贝叶斯框架下，还会为额外参数设置概率分布模型作为先验分布（在方程（9）的荟萃分析模型中为 σ_g、 μ 和 τ，在模型（7）中为 θ_C、 δ、 ψ 和 λ）。参数先验分布的一种常见选择，是采用来自某个分布族的所谓共轭先验，该分布族与数据建模似然分布所属的分布族共轭（Gelman，2006）。为模型参数选择此类共轭先验分布，意味着该参数的后验分布也属于与先验分布相同的分布族。正如 Gelman（2006）、Gelman 等人

（2013）所述，这种做法在解释和计算方面具有优势。根据先验分布所包含的信息量，可将其归类为非信息、弱信息或信息性先验。可以通过先验分布的方差参数来控制这种分类，而且必须结合数据的似然分布来理解（详情见 Gelman 等人（2013））。与数据的影响相比，非信息先验对后验分布的影响极小。其构造方式使其概率密度相对于数据的概率密度而言是平坦的。相较之下，信息性先验分布旨在尽可能准确地体现全部可用的先验知识，并与数据的影响共同对参数的后验分布产生重大影响。弱信息先验分布则是在非信息分布与信息性分布之间折中的选择。它们经过有意设计，比信息性先验分布更平坦。弱信息先验可以通过将后验分布限制在合理的参数范围内，对后验分布起到正则化作用（详情示例见 Gelman 等人（2013）、McElreath（2018））。一般而言，尤其是在估计贝叶斯因子时，优先选择与具体情境相适应的弱信息或信息性先验，而非非信息先验（参见 Gelman（2006）、Betancourt（2017）、Seaman III 等人（2012）、Betancourt（2017）、Schad 等人（2021）、Lemoine（2019））。

2.4.1 方差参数

研究数量较少时，必须特别谨慎地为模型 (10) 中的异质性参数 τ指定先验分布（参见 Gelman（2006年）、Friede 等人（2017*a*））。Friede 等人（2017*a*）建议采用这样的先验分布：其概率质量大部分集中在代表小到大异质性的范围内，而仅有一小部分分配给代表更大异质性的取值。异质性程度的解释取决于所建模参数的尺度。Spiegelhalter 等人（2004年）建议结合数据模型的残差标准差 σ，将异质性划分为以下几类:

异质性	$r := \frac{\tau}{\sigma}$
小	0.0625
中等	0.125
显著	0.25
大	0.5
非常大	1.0

表1：分类 根据 Spiegelhalter 等人的研究，异质性与数据标准差之间的关系 l.（2004年）。

因此，推荐的先验分布属于折叠非中心 t 分布族，其中包括半 t 分布和半正态分布这两种特殊情形。半正态分布 $\mathcal{HN}(\varphi)$具有尺度参数 φ，与均值为零、方差为 φ^2 的标准正态分布相关，其定义为取标准正态分布的绝对值。与半 t 分布相比，半正态分布的尾部较小，因此对极端异质性值赋予的权重较低（参见 Spiegelhalter 等人（2004年））。在本应用示例中，各分布参数的设定反映了表1所分类的异质性假设。

2.4.2 截距参数

截距参数的先验分布选用正态分布；在给定所有其他模型参数的条件下，正态分布是数据正态分布的共轭族。更具体地说，建议将单位信息先验（*UIP*）作为截距先验的起点（参见 Kass 和 Wasserman（1995年）、Röver 等人（2021年））。单位信息先验的信息量（方差）相当于一个典型数据点所包含的信息量。历史数据的单位信息先验设定为以历史数据的均值为中心。严格来说，根据数据中的信息来设定先验分布，就意味着同一信息被使用了两次：一次用于设定先验，另一次则通过观测数据的似然发挥作用；二者都会用于估计后验分布。这与贝叶斯先验分布的概念相矛盾，因为先验分布应代表在看到所分析的（历史）数据之前的先验知识。尽管如此，它仍可作为起点，确保先验分布以某个合理范围为中心。选择一个能使先验分布足够宽、但又不至于宽得不切实际的方差参数，便可使先验仅反映粗略估计，并让数据提供更精确的信息，以细化参数的后验分布，同时发挥正则化作用（McElreath，2018年）。这一点同样可以通过敏感性分析进行检验。单位信息先验也可作为检验贝叶斯假设的参考先验，或用于评估与信息量更大的先验相比，其对后验分布的影响（参见 Kass 和 Wasserman（1995年）、Raftery（1998年）、Neuenschwander 和 Schmidli（2020年）、Li（2021年）、Röver 等人（2021年））。

针对新实验对照组的均值 θ_C 和标准差 σ_C，计算MAP先验。因此，使用 brms 软件包中的 posterior_epred 函数，估计手术组的期望后验预测分布。未手术组的期望后验预测分布用于推导新实验中均值和标准差 θ_C 及 $\sigma_C = \exp(\psi)$ 的参数先验分布，具体说明见下一节。θ_E 和 $\sigma_E = \exp(\psi + \lambda)$ 的先验设定为以与先验 θ_C 和 $\sigma_C = \exp(\psi)$ 相同的值为中心，但采用更大的方差参数，对应于单位信息先验。较大的方差参数会降低先验分布相对于数据的权重。将对照组和实验组参数的先验分布设为以相同值为中心，目的是假定两组的先验均值相等。与此同时，实验组参数先验采用较大方差的原因是：对照组有历史数据可用，因此其先验分布的影响应大于实验组；实验组则没有（直接相关的）历史数据可用。

2.4.3 通过参数分布近似MCMC 抽样值

为将历史数据纳入适当的先验分布，需要用参数分布近似由MCMC抽样值表示的后验预测分布的非参数估计值。Röver 等人（2021年、2022年）介绍了三种将参数分布拟合至MCMC抽样值的一般方法。每种方法的第一步都是指定一个分布族。为历史数据模型中的均值参数选择一个（相对于数据分布而言的）条件共轭正态先验分布，可以确保在其他所有模型参数取固定值的条件下，后验分布属于同一参数族。然而，当其他模型参数不固定时，后验分布不一定是简单的正态分布。对于方程（9）中的正态-正态分层模型（NNHM），Röver（2017年）表明，当 μ 采用正态先验分布、 τ采用半正态先验分布时，新实验中均值参数 μ 和参数 θ^* 的后验预测分布为正态混合分布。因此，还必须考虑更复杂的分布族，以近似后验MCMC抽样值。指定分布族后，估计其参数的一种简单方法是从后验抽样值中获取点估计，例如均值或中位数。然而，将后验抽样值近似为点估计会造成信息减少，并且并不总能反映后验分布的所有重要特征。另一种方法是用边缘分布近似后验抽样值（见Röver 等人（2021年））。第三种方法是选择一个模型族，并通过最大似然（ML）估计、期望最大化（EM）算法或矩匹配（MM）确定其参数。可以使用模型拟合度或预测性能的估计量（例如赤池信息准则（AIC）；Weber等人（2021年））来比较不同的候选分布族。在贝叶斯语境下，正如 Gelman 等人（2014年）和Vehtari 等人（2017年）所述，渡边-赤池信息准则（WAIC）和留一交叉验证（LOO-CV）优于AIC。但由于参数分布采用频率学派方法而非贝叶斯方法进行拟合，因此本文不考虑WAIC和LOO-CV。在本研究中，MAP先验的候选参数分布包括正态分布、正态混合分布和t分布。拟合混合分布时，使用RBesT R包（Weber等人（2021年））中的automixfit函数。该函数使用期望最大化算法拟合混合成分数量逐渐增加的一系列正态混合分布，并根据AIC值选择最佳拟合结果（AIC值会对模型复杂度进行惩罚）。作为比较，使用最大似然估计拟合简单正态分布和非中心t分布。按照第2.4节所述进行先验预测检验，以确保先验分布是合理的。

将预测分布作为新实验中的先验分布进行构建的方法称为荟萃分析预测（MAP）方法（Neuenschwander 等人（2010））， 该方法要求在给定历史数据的情况下，将先验分布明确构建为预测分布（MAP先验）。与这种依次分析历史数据和新数据的方法不同，另一种方法是荟萃分析合并（MAC）方法，即在一次共同分析中同时分析历史数据和新数据。Schmidli 等人（2014）表明，MAP方法在理论上等价于MAC方法。在实践中，MAC方法的优势在于更加直接且更容易实施，因为它只需拟合一个贝叶斯模型，而不必分别为历史数据、根据历史数据推导MAP先验以及分析新数据各拟合一个模型。然而，在设计分析和样本量确定的背景下，MAP方法的优势在于能够更好地判断和控制历史数据对新分析中估计结果的影响。此外，还可以计算*MAP*先验的有效样本量（*ESS*） n_{eff}，以历史数据在异质样本中的MAP先验方差与在同质合并样本中的MAP先验方差之比来量化历史数据的影响及其信息含量。 n_{eff} 可估计在利用历史数据中的信息后，新实验中可以减少的动物数量（前提是新数据与历史数据可交换）（见Neuenschwander 等人（2010年、2020年））。

3 应用示例

作为一项应用实例，针对一项转化性临床前研究中的动物实验开展了基于模拟的设计分析和样本量确定，该实验旨在考察C5aR1受体在小鼠绝经后骨质疏松症状态下对骨质量的作用。通过对小鼠实施卵巢切除术来模拟绝经后骨质疏松症的病理状态。骨质量通过（除其他指标外）μCT扫描中的（无单位）相对骨体积（骨体积与组织体积之比（骨体积/组织体积））进行测量。下文将该实验称为*C5aR1*实验。计划开展的动物实验旨在检验小鼠中C5aR1基因敲除与相对骨体积之间是否存在关联的证据。如果存在这种关联，那么C5aR1基因敲除可能成为潜在处理的基础，并可在人类临床试验中对该处理进行考察。计划中的实验应采用来自C57BL/6J品系的12周龄雌性小鼠的数据；C57BL/6J品系是Jackson实验室研究机构的标准小鼠品系。

3.1 应用示例中的数据和模型

3.1.1 历史数据

提案者可提供先前实验的内部历史数据，用于规划新实验。这些内部数据包括12只卵巢切除小鼠和10只接受假手术的小鼠的数据。不过，提案者的内部历史数据仅代表迄今为止有关小鼠骨质量研究所收集信息的一小部分，因为骨质量是临床前研究中的一个活跃研究课题（例如，Ignatius、Schoengraf、Kreja、Liedert、Recknagel、Kandert、Brenner、Schneider、Lambris 和 Huber-Lang（2011）；Ignatius、Ehrnthaller、Brenner、Kreja、Schoengraf、Lisson、Blakytny、Recknagel、Claes、Gebhard 等（2011）；Mödinger 等（2018）及其中的引文）。纳入其他数据有望提供更多信息，帮助了解小鼠相对骨体积的分布、不同实验组之间及组内的变异性，以及哪些效应量是合理的。然而，现有文献中的大多数数据都不能作为进一步的历史数据使用，因为外部数据中的动物特征（如年龄和动物品系）及其收集时的实验条件与内部历史数据不一致，或者结局采用了不同的方法测量，具有不同的定义或量表。获取对照数据的一种简便且全面的途径是小鼠表型数据库（MPD）（Grubb 等（2004年），Bogue 等（2018年））。这是一个开放获取数据库，收集用于近交系小鼠表征的表型数据。这些数据可用于表征复杂性状之间的相关性，也可用作对照数据或用于表征突变效应（见 Consortium（2007））。MPD 数据采用整洁、标准化的格式，并提供实验方案和数据分析工具。这些数据由全球各地的研究人员提供，由 MPD 工作人员管理；工作人员还会为数据赋予统一的公共本体，例如 Smith 等人（2005）提出的哺乳动物表型本体，从而更容易识别和比较与目标结局相关的数据。

在 MPD 网页上搜索“BV/TV”，会找到 Levy等人（2015年）一项实验中来自 31 个协作交配系（CC）野生型小鼠品系的数据。CC小鼠是由八个遗传差异显著的品系培育而成的重组近交小鼠。它们具有高度的遗传多样性，平均涵盖整个小鼠基因组中 90% 的等位基因多样性（Chesler等人，2008年）。在计划开展的实验中，小鼠将接受卵巢切除术。由于预计卵巢切除术会影响相对骨体积，因此不能单独使用 MPD 中的小鼠数据来估计新实验中相对骨体积均值的后验预测分布。不过，估计野生型小鼠均值的分布，有助于了解设计分析中所考察的合理效应量范围。例如，可以预期，在新实验中，实验组的 C5aR1基因敲除所能达到的最佳效果，或许是完全逆转卵巢切除术的负面影响，使相对骨体积恢复到野生型小鼠的基础水平。此外，纳入外部数据有助于量化不同品系结局的异质性；理想情况下，还可通过比较内部野生型小鼠与 MPD 野生型小鼠的相对骨体积结局，了解内部小鼠品系在多大程度上能够代表整个小鼠基因组。不过，在本例中，没有内部野生型小鼠数据，只有卵巢切除小鼠和假手术小鼠的数据。

外部数据和内部数据见表2。由于假设仅针对雌性动物，因此荟萃分析中排除了外部 MPD 数据集中的所有雄性动物。根据定义，相对骨体积只能取正值。在历史数据中，大多数数值接近零，只有少数数值非常大。对数据进行了对数变换，将结局的取值范围转换为实数，并使其更接近正态分布。

实验	品系	OP	n	$\widehat{\text{E(骨体积/组织体积)}}$	$\widehat{\text{SD(EI)}}$	$\widehat{\text{E(log(EI))}}$	$\widehat{\text{SD(log(EI))}}$
内部	C57BL/6	Ovx	12	1.5	1.10	0.1	0.93
内部	C57BL/6	Sham	10	4.2	1.19	1.4	0.36
MPD	PreCC1061/Tau	None	2	9.2	0.74	2.2	0.08
MPD	PreCC111/Tau	None	9	14.4	3.23	2.6	0.22
MPD	PreCC1156/Tau	None	2	15.8	0.21	2.8	0.01
MPD	PreCC1513/Tau	None	7	29.2	6.47	3.4	0.21
MPD	PreCC188/Tau	None	9	6.9	2.11	1.9	0.29
MPD	PreCC1912/Tau	None	6	12.8	1.17	2.5	0.10
MPD	PreCC2126/Tau	None	6	2.4	2.14	0.6	0.76
MPD	PreCC2156/Tau	None	8	7.7	2.74	2.0	0.34
MPD	PreCC2391/Tau	None	2	2.9	1.76	0.9	0.66
MPD	PreCC2573/Tau	None	3	12.5	9.42	2.4	0.70
MPD	PreCC2680/Tau	None	7	5.8	2.29	1.6	0.53
MPD	PreCC2689/Tau	None	6	6.5	2.24	1.8	0.33
MPD	PreCC2750/Tau	None	5	6.6	3.91	1.8	0.61
MPD	PreCC3348/Tau	None	7	15.1	3.91	2.7	0.24
MPD	PreCC3438/Tau	None	5	7.2	3.46	1.9	0.49
MPD	PreCC3480/Tau	None	3	6.8	7.06	1.4	1.40
MPD	PreCC3912/Tau	None	10	11.3	2.32	2.4	0.20
MPD	PreCC4052/Tau	None	12	13.5	11.47	2.3	0.81
MPD	PreCC4141/Tau	None	6	12.8	5.47	2.5	0.46
MPD	PreCC4438/Tau	None	3	8.5	0.67	2.1	0.08
MPD	PreCC4457/Tau	None	7	14.9	3.83	2.7	0.30
MPD	PreCC519/Tau	None	6	15.6	7.55	2.6	0.48
MPD	PreCC521/Tau	None	7	6.1	1.47	1.8	0.23
MPD	PreCC557/Tau	None	4	5.3	1.72	1.6	0.35
MPD	PreCC611/Tau	None	3	8.7	0.16	2.2	0.02
MPD	PreCC670/Tau	None	3	14.9	3.53	2.7	0.23
MPD	PreCC711/Tau	None	3	12.3	1.01	2.5	0.08
MPD	PreCC72/Tau	None	6	8.1	1.88	2.1	0.23

^ ^ 表2：历史数据中相对骨体积（BV/TV）在原始尺度和对数尺度上的样本量（n）、均值（E）和经验标准差（SD）。估计值根据组别因素实验（内部、MPD）、小鼠品系和手术组（OP，包括卵巢切除术（Ovx）、假手术和未手术）计算。

3.1.2 荟萃分析模型

在贝叶斯框架下，使用 brms 软件包中的 brm 函数（Bürkner (2021)）拟合包含个体数据的层级模型或荟萃分析模型（IP-MA，个体患者荟萃分析）（例如，参见 Lyman & Kuderer (2005)、Michiels 等 (2005)、van Walraven (2010)）。作为比较，使用 nlme 包中的 lme 函数（Pinheiro et al. (2022)）在 R 中拟合频率学派模型，以纳入品系随机效应；限制最大似然（REML）估计值通过最大化对数限制最大似然法得出（参见 Pinheiro & Bates (2006)）。模型中纳入手术组变量 OP 作为预测变量，用以表示小鼠接受了卵巢切除术、假手术还是未接受手术。卵巢切除术和假手术对各自手术组中所有动物的影响可能并不相同，而未手术动物的数据预期具有较小的方差。为了允许卵巢切除组小鼠的残差方差不同于假手术组，并允许其与未手术组的残差方差也不同，拟合了一个分布模型（如 Bürkner (2020) 所述），其中残差方差由其自身特定于手术组的预测变量项建模，类似于方程（7）中的项。理想情况下，鉴于已知骨质量取决于年龄，应将年龄纳入预测变量项。但由于各手术组分别来自不同的年龄区间（未手术动物最年轻，而假手术和卵巢切除动物年龄均较大），纳入年龄变量后，年龄效应和手术组效应无法区分。数据来源的实验（内部数据、MPD）也存在同样的问题。理想情况下，该变量也应作为固定效应或随机效应参数建模。但由于假手术动物和卵巢切除动物均为内部数据，而未手术动物则是来自 MPD 的外部数据，因此这一效应也无法估计。

总之，拟合以下荟萃分析模型：

$$
\begin{aligned}
&y_{ijk} = \alpha + \beta_1 1(i=1) + \beta_2 1(i=2) + \nu_j + \epsilon_{ijk},\\
&\text{with operation group specific residuals}\\
&\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_i),\ \sigma_i = \exp(\eta_{\sigma_i}),\\
&\eta_{\sigma_i} = \psi + \lambda_1 1(i=1) + \lambda_2 1(i=2)\\
&\text{and additional independent variance components}\\
&\nu_j \sim \mathcal{N}(0, \tau_\nu^2)
\end{aligned}
\tag{10}
$$

其中，$k = 1$，……，n_{ij}表示手术组 $i = 0$、1、2（未手术、卵巢切除术（卵巢切除）、假手术）中小鼠的索引，品系为$j = 1$，……，22。先验分布的选择将在下一节中说明。

3.1.3 历史数据的先验分布

在 brms 软件包中，默认指定非信息性或弱信息先验分布。截距（均值的截距，α，以及对数残差标准差的截距，ψ）所采用的默认先验，与适用于整个数据集的总体参数（在频率学派框架中为固定效应，此处为 β_i、λ_i、$i = 1$，2 和 γ）以及用于建模组间异质性的组别特定方差参数（在频率学派框架中为随机效应，此处为 ν_j、$j = 1$，……，22 ）所采用的默认先验不同。

表3总结了模型（10）中各参数手动选取的先验分布，并将其与 brm 函数中的默认先验进行比较。对于截距，按照第2.4节的说明，将单位信息先验（UIP）作为参照。截距 α 表示未手术且年龄为八周的动物的平均相对骨体积（对数尺度）。单位信息先验的信息量相当于单个观测值。因此，这里考虑未手术的历史动物。使用最大似然估计对未手术组拟合正态分布，估计得到的残差标准差为 0。72。为使先验的信息量略低一些，将 α 的先验标准差增至 1。α 的先验均值设为未手术组的估计均值，即 2。这些选择在对数尺度上产生 $[0.1, 4]$ 的95%区间，在原始尺度上则产生 $[1.1, 53]$ 的95%区间。模型（9）中的手术组特异性残差标准差 $\sigma_i, i = 0$，1，2（未手术组、卵巢切除组、假手术组）被建模为正态分布线性预测量 η_{σ_i} 的指数，该预测量以均值 0 为中心。尺度参数为 0。5 时，σ_0 的95%分位数为 2。3；与 brms 中默认的 t_3(0，2。5) 分布相比，这一结果被认为更为合理，因为默认分布会使 σ_0 的95%分位数达到360。

对于方差参数 ν_j、$j = 1$、……、22 的异质性参数 τ_ν（代表不同品系导致的方差），如第 2.4 节所述，选择半正态先验。τ_ν 的尺度参数设为$_{1\ 2}$，以先验 σ_0 的均值为参考尺度，按照表1分类时，这对应较大的异质性。先验上，预计小鼠品系间存在较大的异质性，因为如上所述，这些品系包括来自遗传背景多样的小鼠群体的 CC小鼠。brms 中的总体效应 β_i 和 λ_i、$i = 1$，2 保持默认值，只有在出现收敛问题迹象或先验预测检验表明这些值不合理时才进行修改。这里两种情况都没有出现。

参数	默认	手动
α	t_3(1.7, 2.5)	$\mathcal{N}$(2, 1^2)
ψ	t_3(0, 2.5)	$\mathcal{N}$(0，0。5_2)
β_i	$\mathcal{U}$(−∞，∞)	$\mathcal{U}$(−∞，∞)
λ_i	$\mathcal{U}$(−∞，∞)	$\mathcal{U}$(−∞，∞)
τ_ν	$\mathcal{U}$(−∞，∞)	$\mathcal{HN}$(0，0.5_2)

表3：历史数据对数尺度下贝叶斯荟萃分析模型的先验分布（brms 中的默认设置和手动选择）。

3.1.4 先验预测检验

在先验预测检验中，对包含候选先验分布的模型进行检验；先验预测检验由 Good（1950）提出。在先验预测检验中，根据给定的模型和先验分布模拟大量重复数据，旨在判断模型和参数分布在（生物学上）是否合理。尽管通过审查模型和参数分布的定义便可对其合理性作出判断，但在先验预测检验中最容易观察并评估所有模型组件之间的相互作用，因为生成的数据

反映了所有这些模型组件，并可与研究人员对这一情境下典型数据应呈现何种样貌的先验预期进行比较。为执行先验预测检验，我们根据模型（10）和第 3 节所指定的先验分布生成数据集。对于构建卵巢切除术组和假手术组数据所需的总体效应 $\beta_i, \gamma, \lambda_i, i = 1,2$，选用了不适定的非信息先验 $\mathcal{U}(-\infty, \infty)$。由于无法从该分布中抽样，且为缩小生成图形的范围，先验预测检验仅针对未手术组展示。其他组的假设数据可以通过能够近似 $\mathcal{U}(-\infty, \infty)$分布的先验生成，例如使用方差极大的正态分布 $\mathcal{N}(0, \sigma^2)$。生成 1000 个先验预测数据集。对于每个数据集，使用 R 中的随机数生成函数模拟模型参数，然后根据（10）中描述的模型构建假设数据。对于未手术组中的动物，这相当于

$$\tilde{y}_{r1jk} = \tilde{\alpha}_r + \tilde{\nu}_{r,j[l]} + \tilde{\epsilon}_{r1jk} \tag{11}$$

对于小鼠 $k = 1, \ldots, K$，来自小鼠品系 $j[k] = 1, \ldots, 22$，在假设数据集 $r = 1, \ldots, 1000$ 中。每个模拟数据集中的小鼠数量 K 选定为与历史数据中的小鼠数量（即 $K = 72$）相对应。不过，目标并非选择能够精确反映历史数据分布的先验分布，而是选择弱信息先验分布，使其先验预测分布与实际观测到的历史数据分布相比信息量较少（更平坦），同时排除结合历史数据背景来看似乎高得不合理的数值。其理论依据是：对于大量生成的数据集，模拟数据的经验分布可以近似数据的先验预测分布（参见 Good（1950））。

3.2 应用示例的结果

3.2.1 将贝叶斯正态-正态层级模型拟合至历史数据

图1展示了历史数据中荟萃分析模型先验分布的先验预测检验结果。直方图的分布表明，brms 中的默认先验分布会导致无卵巢切除术组的相对骨体积对数值非常大。10%和90%区间的边界几乎位于 -25 和25处，这对应于原始尺度上约为0和 $7.2 \cdot 10^{10}$ 的值。相比之下，采用弱信息先验时，10%和90%区间的边界位于 -7.5 和7.5处，在原始尺度上仍分别对应相当低和高的值 $5.5 \cdot 10^{-4}$ 和1808。从生物学角度来看，弱信息先验分布似乎更为合理，但仍允许实际观测数据对后验分布产生重大影响。由于模型拟合中未出现收敛问题，后验估计看起来也合理，并且荟萃分析模型并非主要模型，而主要是用于为新实验的分析构建先验，因此不再对荟萃分析模型的先验分布进行进一步微调。不过，也可以考察能够使总体相对骨体积更小的其他先验分布。下文将通过后验预测检验考察所估计荟萃分析模型的合理性。此外，还将检查马尔可夫链蒙特卡洛诊断。诊断结果未显示出发散或高自相关的迹象。

历史数据的荟萃分析模型中总体效应（固定效应）、方差分量（随机效应）的标准差以及残差的估计值，均以均值和95%分位数区间的形式列于表4，并与频率学派分析所得估计值进行比较；后者采用REML估计量，并在正态假设下计算近似置信区间。总体而言，贝叶斯MCMC与频率学派REML方法得出的估计结果相似。残差标准差的估计结果略有较大差异。值得注意的是，对于观测值极少的假手术组和Ovx组，残差标准差的差异更为明显。图2展示了品系效应的森林图。如第2.3.1节所述，层级（随机效应）模型得出的组别特异性估计值会向共同均值（合并效应）收缩，且其方差小于将效应作为固定效应、独立估计时的方差。贝叶斯分析和频率学派分析的点估计值也再次表现出相当高的一致性。

作为贝叶斯工作流的一部分，我们会进行后验预测检验，以确保模型根据原始数据生成合理的数据（详情见 Gelman等（2013））。结果见图3。结合观测数据来看，后验分布似乎是合理的。与相应的先验分布相比（此处仅展示了图3中的未手术组），后验分布的尾部更短，这是因为历史数据提供了更新后的信息。

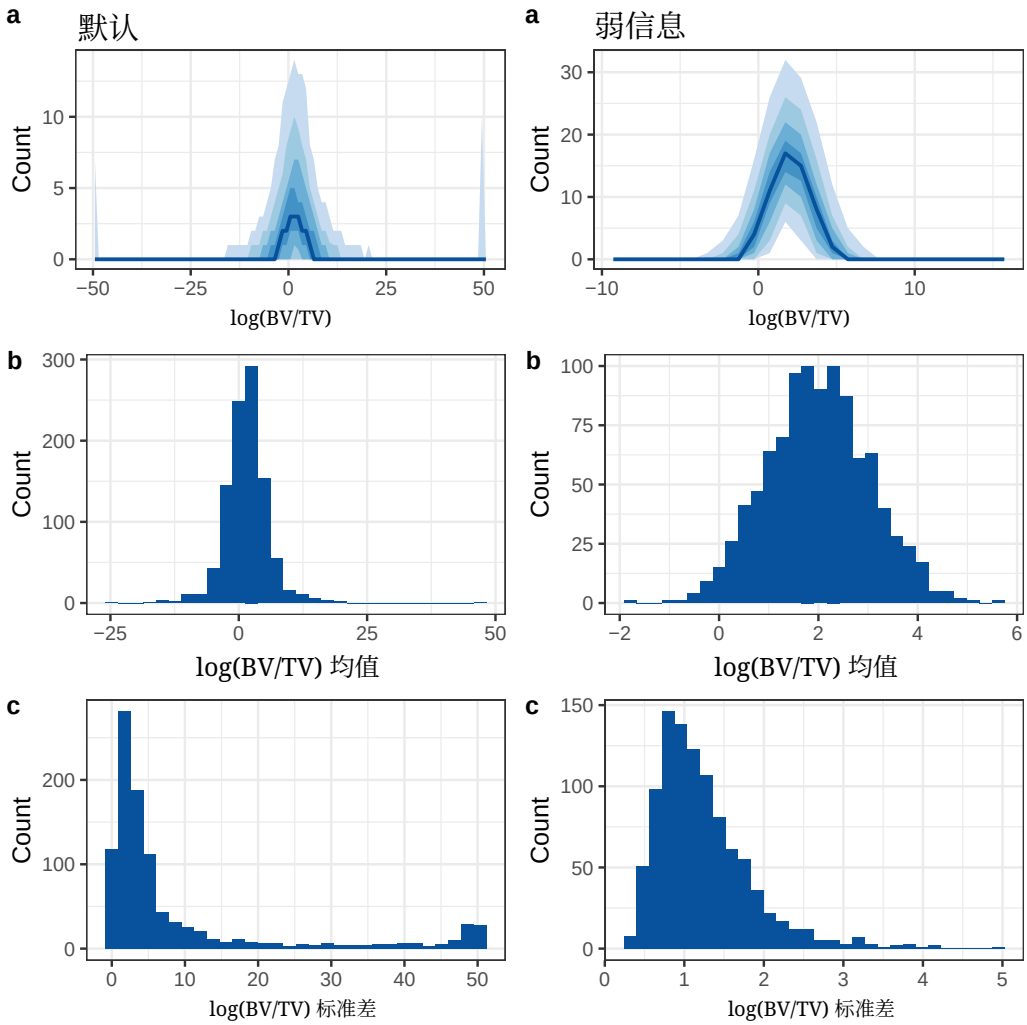


图1：改编自 Schad 等人（2021）的图形化先验预测检验，展示了不同先验分布下未接受手术的动物相对骨体积在对数尺度上的情况。左栏：R 包 brms 中的默认先验分布。右栏：表3中的弱信息先验分布。预测分布基于 1000 个模拟数据集计算得出。a）每个模拟数据集计算出的直方图分布。彩色区域按强度递增的顺序，分别对应所有模拟数据集直方图频数的 10-90、20-80、30-70 和 40-60 百分位区间。区间中间的深色曲线表示所有模拟数据集中位数的分布。b）算术平均数的分布。c）标准差的分布。极端 log(BV/TV) 值 < -50 或 > 50 表示为 -50 和 50，以便展示。

3.2.2 MCMC 抽样值的近似与先验预测分布的定义

表5展示了采用最大似然法近似正态分布，以及采用期望最大化法近似（根据赤池信息准则选出的最佳）正态混合分布并依据赤池信息准则进行选择的结果（如第2.4.3节所述）。两种方法的近似结果看起来相当相似，因此选择采用最大似然法得到的简单正态分布近似作为先验，而不采用更复杂的混合分布，以避免过拟合并简化模型；这是因为前者所需的参数少于包含多个组件的混合分布。

对照组（卵巢切除组）正态分布先验 p（θ_C）的有效样本量，使用Weber等人（2021年）RBesT 包中的 ess 函数计算。计算时需要指定一个参考尺度，作为历史和新对照组动物（组内）残差标准差的估计值。该残差标准差估计值取历史卵巢切除动物残差标准差后验分布估计均值。由此得到的先验有效样本量估计值相当低，为$n_{\mathbf{eff}} = 2$，这表明在本例中，在分析中利用历史数据的信息所带来的益处

变量	贝叶斯	频率学派
截距	2 [1.7,2.2]	2 [1.7,2.3]
Ovx	-1.9 [-3.2,-0.46]	-1.9 [-3.3,-0.51]
Sham	-0.53 [-1.8,0.78]	-0.57 [-1.9,0.74]
品系	0.64 [0.47,0.87]	0.63 [0.45,0.88]
标准差（假手术组）	0.41 [0.254,0.689]	0.36
标准差（卵巢切除组）	1 [0.666,1.62]	0.93
标准差（无处理组）	0.37 [0.301,0.451]	0.35

表4：估计的总体效应（固定效应）、方差分量（随机效应）和组特异性残差的标准差。贝叶斯：马尔可夫链蒙特卡罗模拟的算术平均数及95%分位数区间。频率学派：频率学派REML估计量，其中随机效应的置信区间是在正态假设下的近似置信区间。

变量	分布	方法	组件	权重	E	SD	中位数（95%区间）
μ_C	正态混合	EM	1	1.00	0.10	0.69	0.1 [-1.2,1.5]
μ_C	正态分布	ML	1	1.00	0.10	0.69	0.1 [-1.3,1.5]
σ_C	正态混合	EM	1	0.53	1.10	0.27	1.1 [0.71,1.7]
σ_C	正态混合	EM	2	0.47	0.90	0.15	0.9 [0.65,1.2]
σ_C	正态分布	ML	1	1.00	1.00	0.24	1 [0.64,1.6]
$-\beta_1$	正态混合	EM	1	1.00	1.90	0.70	1.9 [0.49,3.2]
$-\beta_1$	正态分布	ML	1	1.00	1.90	0.70	1.9 [0.53,3.3]

表5：将卵巢切除动物的 MCMC 后验分布近似为参数分布的结果。EM：根据赤池信息准则选出的最佳拟合，即通过期望最大化算法拟合的一系列模型中的正态混合近似。ML：通过最大似然法拟合的正态分布。E 和 SD：相应参数分布的均值和标准差。95%区间：参数分布的0.025和0.975分位数。

新实验的规模较小。如果有更多历史数据可用，就可以为设计分析推导出信息量更丰富的先验分布和模型。讨论中介绍了促进历史数据可用性的方法。

3.2.3 设计分析与样本量确定

模拟新数据中真实 δ 的候选值取自一个范围，该范围从最小值零（对应于无效应，即对照组和实验组的均值平均而言相等）到一个最大值，该最大值对应于拟合卵巢切除动物与未手术动物的预测均值之差的负值所得参数分布的均值（表 5 中的β_1）。参数分布的均值估计为 1.9，标准差估计为 0.7。关于估计的均值标准差，在

卵巢切除组 $\hat{\sigma}_C = \exp(\log(\sigma_C)) = 1$，该估计值对应 Cohen（1988）定义的效应量 $d = \frac{1.9}{1} = 1.9$（根据 Ferguson（2016）和 Sawilowsky（2009）的分类启发式规则，属于非常大到极大的效应量）。此外，还对均值效应量 $\frac{1}{3} \cdot 1.9 \approx 0.6$ 和 $\frac{2}{3} \cdot 1.9 \approx 1.3$ 进行建模；相对于 $\hat{\sigma}_C = 1$，它们分别对应中等和大的效应量。还评估了处理效应为 $\delta = 0$（无效应）的其他设计，以考察 I 类错误和假发现率。此外，还研究了实验组的残差标准差大于对照组的异方差设计（这可能是由处理效应的差异所致）。因此，将系数 λ 设为 log(1.5)，以模拟实验组的标准差为对照组标准差的 1.5 倍的情形（$\sigma_E = \exp(\psi + \lambda) = 1.5\sigma_C$）。样本量候选值采用转化动物实验中常见的样本量，即对照组或实验组各有五只或十只动物。如果实验设计中每组五只动物似乎不足以达到特定的统计目标，而十只又显得过多，则可考察五至十之间更精细的候选样本量。各项设计汇总于表 6。新实验对照组中均值 θ_C 和残差标准差 σ_C 的先验分布与数据分布参数，根据表 5 中正态分布最大似然拟合所得的估计参数设定。θ_E 的先验选用弱信息单位信息先验（UIP），其均值等于 θ_C 的均值，标准差则对应历史去卵巢动物标准差均值的估计值。$\eta_{\sigma_E} = \log(\sigma_E)$ 的先验均值也设为等于 η_{σ_C}的均值。η_{σ_E} 的先验标准差设得高于 η_{σ_C}的标准差，取值为一。采用这些参数后， σ_C 的先验以与

σ_E的先验相同的值为中心，但尾部更宽，因此也允许出现更极端的值。先验预测检验表明，这些先验是合理且具有足够弱信息的，不会压过新数据。

模型	n_E	n_C	μ	δ	log(σ_C)	$\frac{\sigma_E}{\sigma_C}$
1	5	5	0.1	0.0	0.0	1.0
2	10	5	0.1	0.0	0.0	1.0
3	10	10	0.1	0.0	0.0	1.0
4	5	5	0.1	0.6	0.0	1.0
5	10	5	0.1	0.6	0.0	1.0
6	10	10	0.1	0.6	0.0	1.0
7	5	5	0.1	1.3	0.0	1.0
8	10	5	0.1	1.3	0.0	1.0
9	10	10	0.1	1.3	0.0	1.0
10	5	5	0.1	1.9	0.0	1.0
11	10	5	0.1	1.9	0.0	1.0
12	10	10	0.1	1.9	0.0	1.0
13	5	5	0.1	0.0	0.0	1.5
14	10	5	0.1	0.0	0.0	1.5
15	10	10	0.1	0.0	0.0	1.5
16	5	5	0.1	0.6	0.0	1.5
17	10	5	0.1	0.6	0.0	1.5
18	10	10	0.1	0.6	0.0	1.5
19	5	5	0.1	1.3	0.0	1.5
20	10	5	0.1	1.3	0.0	1.5
21	10	10	0.1	1.3	0.0	1.5
22	5	5	0.1	1.9	0.0	1.5
23	10	5	0.1	1.9	0.0	1.5
24	10	10	0.1	1.9	0.0	1.5

表 6：针对对数尺度上的相对骨体积结局进行设计分析所采用的设计（模拟数据的参数和样本量候选值）。

在模型（7）下，使用选定的设计模拟了 10000 个数据集。设计分析结果见图 4、5、6、7 和 8。模拟在巴登-符腾堡高性能计算集群（bwHPC）上运行。借助并行计算、数组作业以及 brms 中的更新功能，计算耗时约 7 小时。图 4 a) 显示了不同实验设计下，Welch 检验的 p 值小于 0.5 的数量、不包含零效应 $\delta = 0$ 的 95% 频率学派置信区间，以及不包含零效应 $\delta = 0$ 的 95% 贝叶斯分位数可信区间；这些结果以总样本量（新实验组和对照组中的动物总数，$n_E + n_C$）为函数绘制。频率学派回归所依据的判定结果与 p 值相当相似，Welch 检验的拒绝比例略高。比较采用不等组均值（$\theta_C \neq \theta_E$）进行模拟的设计中的频率学派曲线和贝叶斯曲线可见，先验分布中的额外信息使贝叶斯模型的拒绝比例高于频率学派模型。在此情境下，检验效能可定义为排除零值的 95% 置信区间（或更准确地说，可信区间）的百分比。在采用相等标准差（$\sigma_C = \sigma_E$）进行模拟的设计中，只有在 $n_E = 10$ 和 $n_C = 5$ 或 $n_C = 10$ 的情况下（当效应较大时，$\delta = 1.9$），以及在贝叶斯模型中 $\delta = 1.3$ 的情况下，检验效能才达到至少 80%。在采用不等标准差进行模拟的设计中，只有在 $n_E = 10$ 和 $n_C = 10$ 的情况下，针对 $\delta = 1.9$，以及在 $n_E = 10$ 和 $n_C = 5$ 的情况下（贝叶斯模型也适用），检验效能才达到至少 80%。图 4 b) 将 a) 中贝叶斯分位数区间的曲线与贝叶斯最高后验密度区间（HDI）得出的曲线进行了比较（关于 HDI 和分位数区间的详情，例如参见 Held & Sabanés Bové (2014)）。

图5展示了将精度作为实验规划的另一项目标时，各种设计的精度。图中显示了随机抽取的一组置信区间或可信区间的宽度，这些区间分别由 Kruschke（2015年）和 Elsey（2021年）提出。使用 HC3 sandwich 估计量，以考虑实验组和对照组的标准差可能不同，许多频率学派置信区间都远宽于分布模型得出的贝叶斯区间。样本量较小时（对照组或实验组中有5只动物），以及残差标准差实际上不相等时 $\frac{\sigma_E}{\sigma_C} = 1.5$，这种情况尤为明显。相比之下，当两组各有10只动物且残差标准差实际上相等时，频率学派区间的宽度与贝叶斯区间更为接近。如果分析设计时关注的目标是达到特定精度，而非检验效能，就必须用目标区间宽度的阈值来定义所需精度。例如，如果目标是以较高的

概率（例如95%）确保标准差相等的设计中所有区间的宽度都不超过1.1，那么在贝叶斯模型中，较大的样本量 $n_E = 10$ 、 $n_C = 5$ 或 $n_C = 10$ 可以达到这一目标，而频率学派模型中的所有样本量都无法达到。

图6显示，对于效应从无到中等的设计，频率学派估计量的均方误差（MSE）平均而言大于贝叶斯模型中的MSE。样本量越大，频率学派模型中的MSE就持续减小。相比之下，在贝叶斯模型中，只有效应量较大的设计，其MSE才会随样本量增加而略微降低。

图7显示了不同设计的平均M型错误率和S型错误率。在大多数设计中，与频率学派模型相比，贝叶斯模型的S型错误率相当相似。不过，在效应量较大且样本量较大的设计中，M型错误率存在显著差异：频率学派模型的M型错误率明显高于贝叶斯模型。对于频率学派模型，所有设计中的M型错误都非常大；而对于贝叶斯模型，随着效应量增大，M型错误会减小，因为先验分布的选择会使效应量的后验分布向零值收缩。这表明，如果采用频率学派分析开展新实验并采用上述任一设计（且新数据确实符合用于生成模拟数据的模型），那么依据设计选择和统计显著性陈述来确定研究方向（即判断置信区间或可信区间是否包含零值），几乎总会导致高估处理效应。所有设计中的S型错误都较小，但在效应较小的模型中约为10%。

对估计贝叶斯因子的分析，可以让人了解支持零假设和备择假设的证据有多充分。不同设计中估计贝叶斯因子的分布见图8。在无效应（$\delta = 0$）或效应较小（$\delta = 0.3$）的设计中，贝叶斯因子的分布有明显的峰值，其大部分概率质量位于1以下，这表明根据对组均值采用等先验分布这一保守选择以及数据提供的证据， H_0 比 H_1更有可能。对于较小的样本量（$n_C = n_E = 5$），分布峰值更接近1；而对于较大的样本量，峰值会进一步向0移动，因为此时相较于样本量较小的情况，有更多证据支持 H_0 。与零效应相比，当 δ 的取值较大时，贝叶斯因子的分布会变得更加平坦，因为此时数据中的信息开始排除先验证据比支持组均值相等这一零假设的倾向。对于非常大的效应（ $\delta = 1.9$），贝叶斯因子的分布峰值位于1以上，表明支持 H_1的证据更多；而对于效应量仅为 $\delta = 1.3$ 的情况，分布峰值仍非常接近1，尤其是在样本量较小以及标准差不相等的情况下。对于具有 $n_E = 10$ 的较大样本量，贝叶斯因子的分布会变得非常平坦，并出现极端值。表7显示，在数据根据带有 $\delta = 0$的零假设进行模拟的设计中，平均而言支持 H_0 的证据更多，而 H_1 的后验模型概率中位数低于50%。设计分析的一个目标可能是找到某个样本量，使 H_1 的后验模型概率的95&分位数区间超过50%。对于效应非常大（ $\delta = 1.9$）且两组标准差相等的设计，当实验组有10只动物、对照组有5只或10只动物时，即可实现这一目标。设计分析的另一个目标可能是找到某个样本量，使贝叶斯因子表明支持 H_1 的证据至少达到中等证据水平的概率不低于80%。对于具有 $\delta = 1.9$ 的设计而言，10只动物的样本量就足够了；这些设计对应于平均而言逆转基因敲除动物中负向的卵巢切除效应。

此外还评估了此处未列出的其他设计。具体而言，我们考察了将 θ_C 的先验标准差降至原来的一半，以使先验信息量更大的效果。然而，这并未产生明显影响。此外，我们还考察了将 δ 的先验标准差增至原来的两倍和二十倍，以使先验信息量更小。这导致假发现率升高，因为先验对后验分布的影响减弱，小数据集中的极端观测值可能导致假阳性结论。此外，即使对于模拟时设定了真实效应量较大或很大的设计，标准差为原来二十倍的先验分布（$0.7 \cdot 20 = 14$）也会使贝叶斯因子的分布非常平坦。值得注意的是，Schad 等人（2022）不建议在贝叶斯因子的情境中使用非信息先验。此外，还在贝叶斯设计下模拟了伪数据，并为所有参数设定了概率分布。其中，平均效应为零的设计曲线（$E(\delta) = 0$）显示，如果效应的标准差较大（例如 $SD(\delta) = 1.2$），不包含零效应的区间比例就会远高于5%。因此，在总样本量为15和20的情况下，以及在两组平均相对骨体积相同但实验组标准差更大的情况下，第一类错误会比频率学派框架通常设定的多得多。由于在贝叶斯模拟框架中，频率学派决策规则和基于贝叶斯区间的决策规则都不是为使此类I类错误率（渐近地）低于5%而设计的，因此这一错误率可能远高于5%。

σ_E/σ_C	δ	n_C	n_E	$p(\mathcal{M}_1\|y)$	BF_{10}	$BF_{10}>3$(%)
1.0	1.9	10.0	10	0.99 [0.73,1]	92 [2.5,2.4e+15]	96.6
1.0	1.9	5.0	10	0.98 [0.62,1]	49 [1.6,1400000]	93.6
1.5	1.9	10.0	10	0.92 [0.47,1]	12 [0.86,1000]	81.8
1.0	1.3	10.0	10	0.91 [0.36,1]	9.6 [0.56,1400]	74.1
1.5	1.9	5.0	10	0.9 [0.43,1]	9.1 [0.75,580]	76.1
1.0	1.9	5.0	5	0.86 [0.45,0.99]	6 [0.82,150]	72.4
1.0	1.3	5.0	10	0.85 [0.34,1]	5.8 [0.51,420]	64.9
1.5	1.3	10.0	10	0.74 [0.31,0.99]	2.9 [0.46,170]	49.3
1.5	1.9	5.0	5	0.71 [0.38,0.98]	2.5 [0.61,43]	43.1
1.5	1.3	5.0	10	0.71 [0.33,0.99]	2.4 [0.49,98]	43.9
1.0	1.3	5.0	5	0.7 [0.33,0.98]	2.3 [0.49,57]	41.7
1.5	1.3	5.0	5	0.58 [0.34,0.96]	1.4 [0.51,23]	24.0
1.0	0.6	10.0	10	0.45 [0.25,0.97]	0.81 [0.32,31]	18.4
1.0	0.6	5.0	10	0.43 [0.27,0.94]	0.77 [0.38,16]	14.5
1.5	0.6	5.0	5	0.43 [0.31,0.89]	0.76 [0.45,7.8]	8.1
1.0	0.6	5.0	5	0.42 [0.29,0.91]	0.74 [0.4,9.8]	10.4
1.5	0.6	5.0	10	0.42 [0.3,0.92]	0.73 [0.42,11]	10.8
1.5	0.6	10.0	10	0.42 [0.27,0.94]	0.71 [0.37,17]	12.5
1.5	0.3	5.0	5	0.41 [0.3,0.84]	0.69 [0.44,5.1]	5.0
1.5	0.0	5.0	5	0.4 [0.3,0.81]	0.67 [0.44,4.3]	3.8
1.0	0.3	5.0	5	0.38 [0.28,0.83]	0.6 [0.39,4.7]	4.7
1.5	0.3	5.0	10	0.37 [0.29,0.83]	0.6 [0.41,4.8]	4.6
1.0	0.0	5.0	5	0.36 [0.28,0.76]	0.57 [0.38,3.2]	2.8
1.5	0.0	5.0	10	0.36 [0.29,0.76]	0.57 [0.41,3.3]	2.8
1.5	0.3	10.0	10	0.36 [0.26,0.87]	0.55 [0.36,6.5]	5.5
1.0	0.3	5.0	10	0.35 [0.27,0.83]	0.53 [0.36,4.9]	4.6
1.5	0.0	10.0	10	0.34 [0.26,0.79]	0.53 [0.36,3.8]	3.3
1.0	0.0	5.0	10	0.33 [0.26,0.72]	0.49 [0.36,2.6]	2.0
1.0	0.3	10.0	10	0.33 [0.23,0.88]	0.48 [0.31,7.3]	6.0
1.0	0.0	10.0	10	0.3 [0.23,0.76]	0.43 [0.3,3.1]	2.6

表7：不同设计下（如左侧四列所示），对模型在 H_1 条件下的证据进行量化的另一种方法。 $p(\mathcal{M}_1|y)$：模型 $\mathcal{M}_1$ 作为 H_1条件下模型的后验概率。BF_{10}：贝叶斯因子 BF_{10} 的中位数及其95%分位数区间。 $BF_{10}>3$：根据Jeffreys（1961）的分类，至少具有中等证据支持 H_1 的百分比。分布基于10000个模拟数据集计算得出-

作为比较，使用频率学派 Welch 检验和 R MESS 软件包中的 power_t_test() 函数，通过求解功效等式进行经典样本量计算（Ekstrøm（2020年））。效应量、组特异性标准差以及对照组和实验组的分配比例（$\frac{n_E}{n_C}$）均采用与表6相同的设计设置进行考察。显著性水平和目标检验效能分别设为常用值0.05和0.8。结果见表8。根据此计算，如果两组的残差标准差相同（$\sigma_E/\sigma_C=1$），且两组的分配比例也相同（$n_E/n_C=1$）（设置13），那么实验组和对照组各需6只动物，便足以在该检验中以至少80%的检验效能检测到至少为1.9的效应（即均值差）。如果计划将大约两倍数量的动物分配至处理组，那么计算结果为对照组 $n_C=5$ 只动物、实验组 $n_E=9$ 只动物，以检测至少这一效应量，且两组标准差相同（设置14）。在效应量相同（ $\delta_{\text{rel}}=1.9$）但实验组标准差较大（$\sigma_E=1.5$）的情况下，若两组分配比例相同，则两组各需9只动物（设置15）；若分配给实验组的动物数量为对照组的两倍，则分别需要6只和11只动物（设置16）。

4 讨论

4.1 摘要

本文讨论了贝叶斯框架下转化动物实验的样本量确定和分析，并将其与零假设显著性检验（NHST）框架下的经典频率学派方法进行了比较。相关考量以一项真实动物实验为例进行说明，该实验研究了C5aR1受体在破骨细胞和成骨细胞中的敲除效应对相对骨体积的影响（C5aR1示例）。样本量的确定取决于对新实验所作的模型假设，以及分析和模型拟合的统计目标。在贝叶斯框架中，这些假设包括模型

设置	δ_{rel}	σ_C	σ_E	分配比例（$\frac{n_E}{n_C}$）	n_C	n_E
1	0.3	1.0	1.0	1	176	176
2	0.3	1.0	1.0	2	132	264
3	0.3	1.0	1.5	1	285	285
4	0.3	1.0	1.5	2	187	373
5	0.6	1.0	1.0	1	45	45
6	0.6	1.0	1.0	2	34	68
7	0.6	1.0	1.5	1	72	72
8	0.6	1.0	1.5	2	48	95
9	1.3	1.0	1.0	1	11	11
10	1.3	1.0	1.0	2	9	17
11	1.3	1.0	1.5	1	17	17
12	1.3	1.0	1.5	2	11	22
13	1.9	1.0	1.0	1	6	6
14	1.9	1.0	1.0	2	5	9
15	1.9	1.0	1.5	1	9	9
16	1.9	1.0	1.5	2	6	11

表8：使用 R MESS 软件包中的 power_t_test() 函数进行经典样本量计算的结果，采用双侧 Welsh 检验、0.8 的检验效能和 0.05 的显著性水平。 δ_{rel}：检验需检测到的最小临床相关效应量。 σ_C, σ_E：新实验中的（估计）标准差。分配比例（$\frac{n_E}{n_C}$）：新实验中小鼠分配至实验组和对照组的比例。 n_C, n_E：新实验所需的样本量。

参数的先验分布。为设定先验分布，我们根据现有历史数据估计了一个贝叶斯荟萃分析模型；这些数据包括新动物实验申请者提供的内部数据，以及来自小鼠表型数据库（MPD）的外部数据。作为比较，我们还拟合了一个频率学派模型，其点估计与前者相当接近。设计分析使用了基于拟合的荟萃分析模型构建的先验分布和伪数据。对新实验中对照组的荟萃分析预测先验有效样本量的估计表明，历史对照数据仅相当于两只动物，这与转化动物实验的样本量规划通常存在很大不确定性的普遍认识相符（参见 Mayer & Muche（2013））。作为设计分析的候选样本量，我们考虑了临床前转化动物实验典型样本量范围的最小值和最大值。处理效应候选范围则根据对拟合历史数据的荟萃分析模型所提供的认识，选取了看似合理的范围。模拟需要大量计算资源，尤其是在评估贝叶斯因子时（本研究使用了巴登-符腾堡高性能计算集群（bwHPC）；采用并行计算、数组作业以及 brms 软件包中的更新功能，对30种设计分别进行10000个模拟数据集的计算，耗时约7小时）。在此示例中，基于效能的样本量计算（这里使用 Welch 检验）表明，当动物在两组间等比例分配、且两组残差标准差均为1时，若要检测均值至少相差1.3，两个组各需11只或更少的动物；若要检测相差1.9或更大的差异，则各需9只或更少的动物。若分配比例不等，且实验组的标准差较大，则分别需要6只和11只动物。然而，M型错误率分析显示，在新实验数据分析得出反对零假设的检验结论时，采用经典频率学派方法进行设计和分析可能会导致效应量被高估的比例很高。若将贝叶斯因子作为目标，并要求模型在 H_1 下的后验概率有95%的概率超过50%这一数值（表示模型在 H_0 和 H_1下的概率相等），那么只有在两组标准差均为1且效应量为1.9的设计中，实验组有10只动物（而对照组有5只动物）才足够。若以80%的概率使贝叶斯因子超过由 Jeffreys（1961）及 Lee与Wagenmakers（2014）界定的、表示 H_1 中等证据的启发式阈值，那么即使标准差不等，在新实验中实验组标准差为1.5、对照组标准差为1的情况下，两组各有10只动物也足够。

与经典频率学派框架相比，贝叶斯荟萃分析预测框架既带来了机遇，也面临着挑战。如果规划新实验的目标是达到一定的统计功效，那么在贝叶斯分析中使用历史数据构建先验预测分布，并不会比采用经典频率学派功效计算降低样本量。不过，基于历史数据进行模拟数据设计分析，并评估额外的模型特征和统计量，有助于更好地呈现和估计模型参数目前存在的不确定性。具体而言，贝叶斯

模型框架允许在分析新实验时，将先验知识（从历史数据中推导得出）正式纳入先验分布。其次，在研究中——尤其是在临床前转化研究这样的早期阶段——几乎总是存在不确定性；与将所有模型参数视为固定参数相比，把它们都建模为随机变量能更好地呈现这些不确定性。第三，采用马尔可夫链蒙特卡洛方法进行估计，也能够处理更复杂的模型，从而或许更准确地表示某些数据。例如，在 C5aR1示例的历史数据和新数据中，针对不同实验组分别对不同的残差标准差进行建模，其估计值似乎比频率学派三明治估计量更精确。对历史数据拟合荟萃分析模型，可提供定量摘要，并可用于为新实验设定先验分布。采用贝叶斯方法，即使只有少数实验或组别，也能通过不同历史实验的均值反映异质性。相比之下，在实验数量少且规模小、数据存在异质性的情况下，频率学派模型较难拟合荟萃分析模型或层级模型（Gelman，2006年；Friede 等人，2017*b*，*a*）。对相似历史实验进行荟萃分析，不仅能提供可用于先验设定的初步估计，也能作为质量控制工具。具体而言，如果某些来自共同均值层级模型的测量值与据称相关的测量值显著不同，那么实验设计或分析中的缺陷可能会显现出来，或者相关结论可能需要加以限定（Walley 等人，2016年）。在对历史数据拟合的荟萃分析模型所估计的更大框架下规划和分析新实验，可能有助于得出更恰当的结论，并推动产生更可重复的结果，从而迈出摆脱动物研究可重复性危机的一步（Ioannidis，2005年；Begley & Ellis，2012年；Begley & Ioannidis，2015年；Loken & Gelman，2017年；Goodman 等人，2016年；Jilka，2016年；Freedman 等人，2015年；Macleod & Mohan，2019年；Voelkl 等人，2020年）。不仅考虑内部数据，也考虑 MPD 等外部数据，可以更全面地了解目标结局的自然变异。这有助于得出更具普适性的结论，而这些结论也可能更容易转化为在人类中的应用，或在其他动物中复现实验结果，正如 Voelkl & Würbel（2016年）及 Voelkl 等人（2020年）所建议的那样。在这一应用示例中，贝叶斯和频率学派荟萃分析模型的点估计相当接近，但贝叶斯方法可以更轻松地将置信区间估计为后验抽样值的分位数。如果对组数较少的分组变量建模异质性，频率学派估计与贝叶斯估计可能会有更大差异。如果数据来源（MPD、内部数据）也纳入建模，就可能出现这种情况。在这种情况下，贝叶斯方法已显示出优越性（Friede 等人，2017*a*，*b*）。此外，通过基于伪数据模拟的设计分析来确定样本量，而不是采用经典的功效不等式来确定样本量，可以解决转化研究中常见的若干问题。具体而言，实验设计和数据中的变异，或许可以用贝叶斯因子这样的连续值来更好地表示，而不是用二元决策规则的结果来表示；因此，直接报告不同设计对应的贝叶斯因子可能更有价值。具体而言，Schad 等（2022）在标准认知实验的背景下指出，许多标准设计所提供的证据不足以支持明确结论，并支持通过不同研究人员和实验室之间共享数据来增加样本量的观点。

关于这些挑战，使用马尔可夫链蒙特卡罗方法拟合贝叶斯模型，至少需要对额外的收敛诊断有基本了解，以确保模型拟合恰当。其他建议步骤包括先验预测检验和后验预测检验，以确保指定的先验分布和后验分布是合理的。贝叶斯框架中通常采用的步骤和决策被称为贝叶斯工作流，Gelman、Vehtari、Simpson、Margossian、Carpenter、Yao、Kennedy、Gabry、Bürkner 和 Modrák（2020）对此进行了描述，而这一工作流可能相当复杂。特别是，与频率学派推断相比，贝叶斯推断通常有更多决定因素，因为需要为所有模型参数指定先验分布，并设置 MCMC 参数。在贝叶斯方法应用于设计分析框架时，情况更为复杂，因为不同的设计选项会带来额外的决定因素。决定因素如此之多，难以理解单独改变某个决定因素对后验推断的影响；而考察所有决定因素组合也很快会变得难以理解。此外，关于贝叶斯框架下应采用哪种样本量确定程序（例如，是否应使用决策函数和阈值，以及使用哪些决策函数和阈值），目前文献中尚未达成共识。如果只有少量规模较小的既往实验可供参考，就必须特别关注先验模型及其超参数所依据的假设。尤其重要的是，为异质性参数建立合理的模型，以恰当地反映所关注背景总体中的变异，避免得出过度自信的结论，而这些结论可能仅适用于单项实验中的标准化动物。关于在设计分析和样本量确定中使用贝叶斯因子，首先面临的挑战是如何定义阈值。其次，正如 Schad 等人（2021）所示，贝叶斯因子对先验分布的选择高度敏感；而通常使用桥接抽样或 Savage-Dickey 方法进行估计时，需要大量马尔可夫链蒙特卡罗迭代才能获得稳定结果（Gronau 等人，2020），最好还要重复估计数次。因此，贝叶斯因子的估计在计算上也颇具挑战。最后，应通过基于模拟的校准（SBC）检查贝叶斯因子估计值可能存在的偏倚（Schad 等人，2021）。这些贝叶斯因子工作流程又进一步加重了基于模拟的设计分析本已很高的手动和计算负担；这也可能成为转化研究应用中的障碍，因为研究人员的资源往往相当有限。关于

荟萃分析，面临的挑战包括：由于临床前研究每年的发表数量非常庞大（Bannach-Brown 等人，2021），很难找到相关的历史信息；已发表的估计值往往也受到偏倚影响，例如发表偏倚（Sena 等人，2010；ter Riet 等人，2012；Conradi 和 Joffe，2017；夏里特健康研究院（BIH））。此外，相关文献往往缺乏组织，结局也没有采用统一术语，因此难以比较不同实验的结果（Smith 等人，2005）。这些情况使转化动物实验的研究人员目前难以理解并正确应用贝叶斯方法，也使样本量确定在这一背景下过于繁琐，难以实际应用；要解决这一问题，需要开发便于使用的常规流程和应用工具。

4.2 扩展

本文介绍了若干扩展方法，可用于规划和分析转化动物实验，这些方法采用贝叶斯荟萃分析预测方法和模拟数据设计分析。在本研究中，我们使用贝叶斯因子的分布来直观呈现不同模型下零假设与备择假设的证据比。为了将该分布转化为决策函数，我们基于 Jeffreys 的分类方案 Jeffreys（1961）设定了一个启发式阈值，或者检查 H_1 ($p(\mathcal{M}_1|y)$) 下模型的 95% 后验模型概率是否超过 50%（这代表模型在 H_1 下与在 H_0 下的概率相等）。设定贝叶斯因子阈值的一种更系统的方法，是采用效用函数的定义，如 Schad 等人（2021）以及 Schönbrodt 和 Wagenmakers（2018）所示。贝叶斯因子比较两个相对模型（此处为 H_0 下的模型与 H_1 下的模型）的“样本外”预测性能。判断数据中是否存在证据表明效应 δ 与零假设所述的情况不同，另一种方法是通过考察后验预测来比较样本外预测性能。衡量模型样本外预测性能的一种常见效用函数是期望对数逐点预测密度（ELPD）（详情参见 Gelman 等人（2014）；关于贝叶斯框架下的实际估计方法，参见 Vehtari 等人（2017））。

在本研究中，我们对历史数据拟合了一个贝叶斯荟萃分析模型，旨在获得先验分布并确定合理的设计分析方案。除了贝叶斯荟萃分析模型之外，也可以使用频率学派荟萃分析模型的估计值来设定参数分布，从而定义用于模拟数据设计分析的先验分布和抽样分布。例如，Schad 等（2022）在考察不同假设下贝叶斯因子行为的模拟研究中采用了这种方法。在本研究中，由于缺少必要的数据，无法将年龄以及历史数据中的实验/实验室效应纳入分析。如果可获得的历史数据很少，则难以检验与正态分布相对应的假设，此时非参数模型可能更为合适（Konietschke 等人，2021）。Burr 和 Doss（2005）提出了一种贝叶斯半参数荟萃分析模型，该模型通过狄利克雷过程先验的一个变体对实验特异性效应进行建模，并在 R 包 bspmma 中实现了该模型（Burr，2012）。如果正态假设存在疑问，可以使用该模型；或者，也可以将其与参数模型进行比较，并利用经验贝叶斯方法判断哪种模型更为合适（Burr，2012）。目前，人们正在开展促进系统综述和荟萃分析在动物研究中应用的工作。这类工作的例子包括爱丁堡的实验研究动物数据荟萃分析与综述协作项目（CAMARADES），该项目提供了开展动物试验系统综述和荟萃分析的方法学建议与工具（2021）；以及在线综述平台系统综述平台（SyRF）（Bahor 等人，2021）。此外，人们还在努力将动物数据收集到共同的大型数据库中（Eppig 等人，2005；Blake 等人，2006；Consortium，2007；Hancock 等人，2008；汉诺威，2022；Pognan 等人，2021）。与此同时，动物试验强制实施（预）注册也取得了进展，这可能增加公开可用数据的数量和质量，并减少偏倚（Chamuleau 等人，2018；Bert 等人，2019；Heinl 等人，2019；Baker，2019；van der Naald 等人，2020）。随着这些进展，未来拟合更好的历史证据荟萃分析模型是可行的。设计分析中的模拟伪数据代表了对新实验数据的假设。这些假设基于历史数据，而历史数据已在荟萃分析模型中得到汇总。在先验分布与未来数据的真实分布相匹配这一理想情况下，先验仅代表额外信息，使后验估计更加精确。然而，如果数据的先验预测分布将其大部分概率质量置于根据观测数据的经验分布来看不太可能出现的区域，就会出现所谓的 先验与数据冲突。当先验与数据冲突导致基于后验抽样值的推断失效时，其影响可能成为一个问题（例如参见 Box（1980）、Evans 和 Moshonov（2006））。为了考察所选先验分布中先验与数据冲突对后验推断的影响，可以模拟偏离先验预测分布的数据。在使用贝叶斯模型进行设计分析时，还需要进一步开展工作，以更好地理解和可视化所有影响后验推断的决定性因素。近期，Kallioinen 等人（2021）开发了 R 包 priorsense，旨在帮助研究先验相对于观测数据的影响。此类软件包可能有助于更好地了解所选先验分布如何影响所需样本量和检验决策。如果先验与数据

冲突被认为是所选先验分布存在的问题，则可以按照 Schmidli 等人（2014）的建议，通过向相应分布中加入信息量较低的混合成分来增强先验分布的稳健性。
除了使用基于数据的先验来设计和分析新实验之外，也可以通过先验 elicitation 来选择先验（Garthwaite 等人（2005）；O'Hagan 等人（2006））。在先验 elicitation 方法中，分析人员不会直接指定先验（例如像本文这样基于历史数据指定），而是由领域专家描述目标结局的特性，再由分析人员将这些描述形式化为先验分布。先验 elicitation 可能是本文所用基于数据的先验的一种颇具吸引力的替代方案，尤其是在没有可用历史数据的情况下。不过，就目前而言，技术、实践和社会方面的挑战阻碍了先验 elicitation 的应用（Mikkola 等人（2021））。

致谢

作者感谢来自德国乌尔姆骨科研究与生物力学研究所的 Melanie Haffner-Luntzer 提供动物数据，并就动物表型数据库的使用进行讨论。此外，作者感谢巴登-符腾堡州通过 bwHPC 提供支持。

利益冲突

所有作者均声明不存在任何潜在利益冲突。

资助

本研究由 德国联邦教育与研究部（BMBF）资助，资助编号为 031L0233 .

参考文献

Bahor, Z., Liao, J., Currie, G., Ayder, C., Macleod, M., McCann, S. K., Bannach-Brown, A., Wever, K., Soliman, N., Wang, Q. 等（2021年），'在线系统综述平台的开发与应用：Camarades系统综述设施（SYRF）的早期发展'，*BMJ Open Science* **5**(1), e100103. Baker, M.（2019年），'动物注册库旨在减少偏倚'，*Nature* **573**(7773), 297–299. Bannach-Brown, A., Hair, K., Bahor, Z., Soliman, N., Macleod, M. & Liao, J.（2021年），'临床前元研究中的技术进展'，*BMJ Open Science* **5**(1), e100131. Bartoš, F., Gronau, Q. F., Timmers, B., Otte, W. M., Ly, A. & Wagenmakers, E.-J.（2021年），'医学中的贝叶斯模型平均荟萃分析'，《医学统计学》**40**(30), 6743–6761. Beckers, J., Wurst, W. & De Angelis, M. H.（2009年），'迈向更好的小鼠模型：增强基因型、系统性表型分析与环境型建模'，*Nature Reviews Genetics***10**(6), 371–380. Begley, C. G. & Ellis, L. M.（2012年），'提高临床前癌症研究的标准'，*Nature* **483**(7391), 531–533. Begley, C. G. & Ioannidis, J. P.（2015年），'科学的可重复性：提高基础与临床前研究的标准'，*Circulation research***116**(1), 116–126. Bennett, C. H.（1976年），'根据蒙特卡洛数据高效估计自由能差'，*Journal of ComputationalPhysics***22**(2), 245–268. Bert, B., Heinl, C., Chmielewska, J., Schwarz, F., Grune, B., Hensel, A., Greiner, M. & Schönfelder, G.（2019年），'改进动物研究：动物研究注册库'，*PLoS biology* **17**(10), e3000463. Betancourt, M.（2017年），'弱信息先验的形状如何影响推断'，*Stan*用户指南。*3*月**17**. Blake, J. A., Eppig, J. T., Bult, C. J., Kadin, J. A. & Richardson, J. E.（2006年），'小鼠基因组数据库（MGD）：更新与增强'，*Nucleic acids research***34**(suppl_1), D562–D567. Bogue, M. A., Grubb, S. C., Walton, D. O., Philip, V. M., Kolishovski, G., Stearns, T., Dunn, M. H., Skelly, D. A., Kadakkuzha, B., TeHennepe, G. 等（2018年），'小鼠表型数据库：用于整理实验室小鼠实证表型数据的综合数据库与分析套件'，*Nucleic acids research* **46**(D1), D843–D850. Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R., Joels 等（2019年），'Repair：动物实验的统计功效解决方案'，*BioRxiv* 第864652页。

Box, G. E.（1980年），‘科学建模与稳健性中的抽样和贝叶斯推断’，英国皇家统计学会学报：A辑（综合）143(4), 383–404. Bundesministerium der Justitz（2010年），‘关于保护用于实验或其他科学目的动物的条例（动物保护实验动物条例——tierschversv）’，https://www.gesetze-im-internet.de/tierschversv/BJNR312600013.html. Bürkner, P.（2020年），‘使用brms估计分布模型’。网址：https://cran.r-project.org/web/packages/brms/vignettes/brms_distreg.html Burr, D.（2012年），‘bspmma：用于荟萃分析的贝叶斯半参数模型R软件包’，《统计软件杂志》50(4), 1–23. 网址：http://www.jstatsoft.org/v50/i04/ Burr, D. & Doss, H.（2005年），‘随机效应荟萃分析的贝叶斯半参数模型’，《美国统计学会杂志》100(469), 242–251. Bürkner, P.-C.（2021年），‘使用brms和Stan在R中进行贝叶斯项目反应建模’，《统计软件杂志》100(5), 1–54. Chamuleau, S. A., Van Der Naald, M., Climent, A. M., Kraaijeveld, A. O., Wever, K. E., Duncker, D. J., Fernández- Avilés, F. & Bolli, R.（2018年），‘心血管修复的转化研究：呼吁范式转变’，Circulation research 122(2), 310–318. Chesler, E. J., Miller, D. R., Branstetter, L. R., Galloway, L. D., Jackson, B. L., Philip, V. M., Voy, B. H., Culiat, C. T., Threadgill, D. W., Williams, R. W. 等（2008年），‘橡树岭国家实验室的协作交配系：开发系统遗传学的强大资源’，《哺乳动物基因组》19(6), 382–389. Cohen, J.（1988年），《行为科学的统计功效分析》，劳特利奇。Conradi, U. & Joffe, A. R.（2017年），‘在2008年重症医学会会议上报告的动物研究中的发表偏倚’，《BMC研究札记》10(1), 1–11. Consortium, T. M. P. D. I.（2007年），‘小鼠表型组数据资源的整合’，《哺乳动物基因组》18(3), 157–163. 网址：https://doi.org/10.1007/s00335-007-9004-x Dickey, J. M. & Lientz, B.（1970年），‘加权似然比、关于概率的尖锐假设以及马尔可夫链的阶数’，《数学统计年刊》页码214–226。Ekstrøm, C. T.（2020年），MESS：杂项深奥统计脚本。R软件包版本0.5.7。网址：https://CRAN.R-project.org/package=MESS Elsey, J.（2021年），‘使用贝叶斯估计进行强效序贯设计：使用brms、tidyverse和furrr的功效分析教程’。网址：psyarxiv.com/kt4pz Eppig, J. T., Group, M. G. D., Bult, C. J., Group, M. G. D., Kadin, J. A., Group, M. G. D., Richardson, J. E., Group, M. G. D., Blake, J. A. & Group, M. G. D.（2005年），‘小鼠基因组数据库（MGD）：从基因到小鼠——一个服务于小鼠生物学的社区资源’，《核酸研究》33(suppl_1), D471–D475. Evans, M. & Moshonov, H.（2006年），‘检查先验与数据冲突’，贝叶斯分析1(4), 893–914. Ferguson, C. J.（2016年），‘效应量入门：临床医生与研究人员指南。’美国心理学会。Freedman, L. P., Cockburn, I. M. & Simcoe, T. S.（2015年），‘临床前研究中可重复性的经济学’，PLoS Biol 13(6), e1002165. Friede, T., Röver, C., Wandel, S. & Neuenschwander, B.（2017a年），‘罕见病中少数小型研究的荟萃分析’，Research Synthesis Methods 8(1), 79–91. Friede, T., Röver, C., Wandel, S. & Neuenschwander, B.（2017b年），‘存在异质性时两项研究的荟萃分析及其在罕见病中的应用’，Biometrical journal 59(4), 658–671. Garthwaite, P. H., Kadane, J. B. & O’Hagan, A.（2005年），‘获取概率分布的统计方法’，Journal of the American statistical Association 100(470), 680–701. Gelman, A.（2006年），‘层级模型中方差参数的先验分布（对Browne和Draper文章的评论）’，贝叶斯分析1(3), 515–534. Gelman, A. & Carlin, J.（2014年），‘超越功效计算：评估S类（符号）和M类（幅度）错误’，《心理科学展望》9(6), 641–651。

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. 与 Rubin, D. B.（2013），《贝叶斯数据分析》，CRC出版社。Gelman, A., Hill, J. 与 Vehtari, A.（2020），《设计与样本量决策》，剑桥大学出版社，第291–312页。Gelman, A., Hwang, J. 与 Vehtari, A.（2014），“理解贝叶斯模型的预测信息准则”，《统计与计算》24(6)，997–1016。Gelman, A. 与 Vákár, M.（2019），“揭穿假手术：一种利用含噪对照数据进行自适应调整的贝叶斯模型”，arXiv预印本 arXiv:1905.09693。Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C. 与 Modrák, M.（2020），“贝叶斯工作流”，arXiv预印本 arXiv:2011.01808。Good, I. J.（1950），《概率与证据权衡》，C. Griffin，伦敦。Goodman, S. N., Fanelli, D. 与 Ioannidis, J. P.（2016），“研究可重复性意味着什么？”，《科学转化医学》8(341)，341ps12–341ps12。Gronau, Q. F., Singmann, H. 与 Wagenmakers, E.-J.（2020），“bridgesampling：用于估计归一化常数的R软件包”，《统计软件杂志》92(10)，1–29。Grubb, S. C., Churchill, G. A. 与 Bogue, M. A.（2004），“近交系小鼠特征数据库”，《生物信息学》20(16)，2857–2859。Hancock, J. M., Schofield, P. N., Chandras, C., Zouberakis, M., Aidinis, V., Smedley, D., Rosenthal, N. 与 Schughart, K.（2008），Casimir：国际小鼠信息学资源的协调与可持续性，载于“2008年第8届IEEE生物信息学与生物工程国际会议”，IEEE，第1–6页。Hannover, F. I.（2022），“RITA工业毒理学动物数据登记库”。URL：https://reni.item.fraunhofer.de/reni/public/rita Hartung, J. 与 Knapp, G.（2001），“针对具有二分类结局的对照临床试验的荟萃分析改进方法”，《医学统计学》20(24)，3875–3889。Hedges, L. V. 与 Olkin, I.（1985），《荟萃分析的统计方法》，学术出版社。Heinl, C., Chmielewska, J., Olevska, A., Grune, B., Schönfelder, G. 与 Bert, B.（2019），“重新思考科学激励体系：动物研究登记库：使用动物进行实验预注册可显著提高生物医学研究的透明度和可靠性，并改善动物福利”，《EMBO报告》21(1)，e49709。Held, L. 与 Sabanés Bové, D.（2014），《应用统计推断》，施普林格，海德堡、纽约、多德雷赫特、伦敦。Higgins, J. P. 与 Green, S.（2011），《干预措施系统综述考克兰手册》，John Wiley & Sons。Ignatius, A., Ehrnthaller, C., Brenner, R. E., Kreja, L., Schoengraf, P., Lisson, P., Blakytny, R., Recknagel, S., Claes, L., Gebhard, F. 等（2011），“过敏毒素受体C5aR存在于大鼠骨折愈合过程中，并在体外介导成骨细胞迁移”，《创伤杂志》71(4)，952。Ignatius, A., Schoengraf, P., Kreja, L., Liedert, A., Recknagel, S., Kandert, S., Brenner, R. E., Schneider, M., Lambris, J. D. 与 Huber-Lang, M.（2011），“补体C3a和C5a与IL-1β协同调节成骨细胞的破骨细胞形成和炎症反应”，《细胞生物化学杂志》112(9)，2594–2605。Ioannidis, J. P.（2005），“为什么大多数已发表的研究发现都是错误的”，《公共科学图书馆·医学》2(8)，e124。Jaynes, E. T. 与 Kempthorne, O.（1976），“置信区间与贝叶斯区间之比较”，载于《概率论、统计推断与科学统计理论基础》，施普林格，第175–257页。Jeffreys, H.（1961），《概率论》，牛津大学出版社。Jilka, R. L.（2016），“动物研究走向可重复性的道路”，《骨与矿物质研究杂志》31(7)，1317–1319。Kallioinen, N., Paananen, T., Bürkner, P.-C. 与 Vehtari, A.（2021），“通过幂尺度法检测和诊断先验与似然敏感性”，arXiv预印本 arXiv:2107.14054。Kass, R. E. 与 Wasserman, L.（1995），“嵌套假设的参考贝叶斯检验及其与施瓦茨准则的关系”，《美国统计学会杂志》90(431)，928–934。Keenan, C., Elmore, S., Francke-Carroll, S., Kemp, R., Kerlin, R., Peddada, S., Pletcher, J., Rinke, M., Schmidt, S. P., Taylor, I. 等（2009），“使用啮齿动物增生性病变历史对照数据的最佳实践”，《毒理病理学》37(5)，679–693。Konietschke, F., Schwab, K. 与 Pauly, M.（2021），“小样本量：高维数据分析中的大数据问题”，《医学研究统计方法》30(3)，687–701。

Kramer, M. 与 Font, E.（2017），“在动物实验中减少样本量：历史对照及相关策略”，《生物学评论》92(1)，431–445。Kruschke, J.（2015），《贝叶斯数据分析：使用R、JAGS和Stan的教程》，学术出版社。Lee, M. D. 与 Wagenmakers, E.-J.（2014），《贝叶斯认知建模：实用课程》，剑桥大学出版社。Lehmann, E. L. 与 Romano, J. P.（2006），《统计假设检验》，Springer Science & Business Media。Lehmann, E. L. 与 Romano, J. P.（2022），《一致最强检验》，施普林格国际出版，沙姆，第61–124页。Lemoine, N. P.（2019），“超越非信息先验：为何以及如何在贝叶斯分析中选择弱信息先验”，《奥伊科斯》128(7)，912–928。Levy, R., Mott, R. F., Iraqi, F. A. 与 Gabet, Y.（2015），“遗传关联研究中的协作交配小鼠揭示骨微结构的新候选基因”，《BMC基因组学》16(1)，1–14。Li, Y.（2021），“正态终点的RBest”，https://cran.r-project.org/web/packages/RBesT/vignettes/ introduction_normal.html。访问日期：03.02.2022。Loken, E. 与 Gelman, A.（2017），“测量误差与重复性危机”，《科学》355(6325)，584–585。Long, J. S. 与 Ervin, L. H.（2000），“在线性回归模型中使用异方差稳健标准误”，《The American Statistician》54(3)，217–224。Lyman, G. H. 与 Kuderer, N. M.（2005），“基于汇总数据的荟萃分析的优势与局限”，《BMC医学研究方法学》5(1)，1–7。Macleod, M. 与 Mohan, S.（2019），“动物研究的可重复性与严谨性”，《ILAR期刊》60(1)，17–23。Maddatu, T. P., Grubb, S. C., Bult, C. J. 与 Bogue, M. A.（2012），“小鼠表型数据库（MPD）”，《核酸研究》40(D1)，D887–D894。Makowski, D., Ben-Shachar, M. S. 与 Lüdecke, D.（2019），“bayestestR：在贝叶斯框架内描述效应及其不确定性、存在性与显著性”，《开放源代码软件期刊》4(40)，1541。URL：https://joss.theoj.org/papers/10.21105/joss.01541 Mayer, B., Allgoewer, A. 与 Muche, R.（2018），“根据3R原则，临床前研究动物实验生物统计样本量计算的基本标准”，《柏林与慕尼黑兽医周刊》131(7-8)，272–278。Mayer, B. 与 Muche, R.（2013），“医学基础研究动物实验中正式样本量规划的有限解释力”，《兽医实践K辑：小动物/家养动物》41(06)，367–374。McElreath, R.（2018），《统计学再思考：一门以R和Stan为例的贝叶斯课程》，查普曼与霍尔/CRC出版社。McEntyre, J., Sarkans, U. 与 Brazma, A.（2015），“BioStudies数据库”，《分子系统生物学》11(12)，847。Michiels, S., Baujat, B., Mahé, C., Sargent, D. 与 Pignon, J.（2005），“随机效应生存模型有助于更好地理解个体患者数据荟萃分析中的异质性”，《临床流行病学杂志》58(3)，238–245。Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C. 与 Klami, A.（2021），“先验知识获取：过去、现在与未来”。URL：https://arxiv.org/abs/2112.01380 Mödinger, Y., Rapp, A., Pazmandi, J., Vikman, A., Holzmann, K., Haffner-Luntzer, M., Huber-Lang, M. 与 Ignatius, A.（2018），“C5aR1与成骨细胞中的Tlr 2相互作用，并刺激诱导破骨细胞的趋化因子Cxcl 10”，《细胞与分子医学杂志》22(12)，6002–6014。Neuenschwander, B., Capkun-Niggli, G., Branson, M. 与 Spiegelhalter, D. J.（2010），“总结临床试验中对照组的历史信息”，《临床试验》7(1)，5–18。Neuenschwander, B. 与 Schmidli, H.（2020），“历史数据的使用”，《药物研究中的贝叶斯方法》。Neuenschwander, B., Wandel, S., Roychoudhury, S. 与 Bailey, S.（2016），“适用于具有多个分层的早期临床试验的稳健可交换性设计”，《药物统计学》15(2)，123–134。Neuenschwander, B., Weber, S., Schmidli, H. 与 O’Hagan, A.（2020），“预测一致的先验有效样本量”，《生物统计学》76(2)，578–587。Novick, S. 与 Zhang, T.（2021），“确保临床前药物发现可重复性的均值比较与功效计算”，《医学统计学》40(6)，1414–1428。

爱丁堡大学（2021），《CAMARADES——实验研究动物数据荟萃分析与综述协作项目》。网址：https://www.ed.ac.uk/clinical-brain-sciences/research/camarades 夏里特健康研究所（BIH）（2021），《动物研究中发表偏倚的程度、预测因素及管理（Embarc）》。 [在线；截至11月3日 2021]。网址：https://www.bihealth.org/de/translation/innovationstreiber/quest-center/projekte/translationale-bioethik/embarc O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. & Rakow, T.（2006），《不确定的判断： eliciting experts' probabilities》，John Wiley & Sons。Pinheiro, J. & Bates, D.（2006），《S和S-PLUS中的混合效应模型》，施普林格科学与商业媒体。Pinheiro, J., Bates, D. & R Core Team（2022），nlme：线性和非线性混合效应模型。R软件包版本3.1-157。网址：https://CRAN.R-project.org/package=nlme Pognan, F., Steger-Hartmann, T., Díaz, C., Blomberg, N., Bringezu, F., Briggs, K., Callegaro, G., Capella-Gutierrez, S., Centeno, E., Corvi, J. 等（2021），《ETRANSAFE项目：通过整合性知识管理开展转化安全性评估：成果与展望》，Pharmaceuticals 14(3), 237。Pullenayegum, E. M.（2011），《二分类结局荟萃分析中研究间异质性的知情参考先验》，《医学统计学》30(26), 3082–3094。R Core Team（2021），《R：统计计算语言与环境》，R 统计计算基金会。网址：https://www.R-project.org/ Raftery, A. E.（1998），《贝叶斯因子与贝叶斯信息准则：对"模型选择中贝叶斯信息准则的批判"的评论》，Sociological Methods & Research 27(3), 411–427。Rhodes, K. M., Turner, R. M. & Higgins, J. P.（2015），《为连续结局数据荟萃分析中的异质性程度构建预测分布》，《临床流行病学杂志》68(1), 52–60。Röver, C.（2017），《使用 bayesmeta R软件包进行贝叶斯随机效应荟萃分析》，arXiv预印本 arXiv:1711.08683。Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S. & Friede, T.（2021），《关于贝叶斯随机效应荟萃分析中异质性参数的弱信息先验分布》，Research Synthesis Methods 12(4), 448–474。Röver, C. & Friede, T.（2021），《收缩估计中外部数据权重的界限》，Biometrical Journal 63(5), 1131–1143。Röver, C., Sturtz, S., Lilienthal, J., Bender, R. & Friede, T.（2022），《总结贝叶斯随机效应荟萃分析中研究间异质性的经验信息》，arXiv预印本 arXiv:2202.12538。Russel, W. & Burch, L.（1959），《人道实验技术原则》，Methuen。Satterthwaite, F. E.（1941），《方差综合》，Psychometrika 6(5), 309–316。Sawilowsky, S. S.（2009），《效应量的新经验法则》，Journal of Modern Applied Statistical Methods 8(2), 26。Schad, D. J., Betancourt, M. & Vasishth, S.（2021），《迈向认知科学中有原则的贝叶斯工作流》，Psychological Methods 26(1), 103。Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M. & Vasishth, S.（2022），《稳健使用贝叶斯因子的工作流技术》，Psychological Methods。Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. & Neuenschwander, B.（2014），《在具有历史对照信息的临床试验中采用稳健的荟萃分析预测先验》，《生物统计学》70(4), 1023–1032。Schönbrodt, F. D. & Wagenmakers, E.-J.（2018），《贝叶斯因子设计分析：规划有说服力的证据》，Psychonomic Bulletin & Review 25(1), 128–142。Seaman III, J. W., Seaman Jr, J. W. & Stamey, J. D.（2012），《指定非信息先验的隐患》，The American Statistician 66(2), 77–84。Sena, E. S., Van Der Worp, H. B., Bath, P. M., Howells, D. W. & Macleod, M. R.（2010），《动物卒中研究报告中的发表偏倚导致疗效被严重夸大》，PLoS Biology 8(3), e1000344。Smith, C. L., Goldsmith, C.-A. W. & Eppig, J. T.（2005），《哺乳动物表型本体作为注释、分析和比较表型信息的工具》，Genome Biology 6(1), 1–9。

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P.（2004），临床试验与卫生保健评估的贝叶斯方法，第13卷，John Wiley & Sons。Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D. & Wagenmakers, E.-J.（2019），《使用信息性先验的贝叶斯因子设计分析教程》，《行为研究方法》**51**(3), 1042–1058。Steger-Hartmann, T., Kreuchwig, A., Vaas, L., Wichard, J., Bringezu, F., Amberg, A., Muster, W., Pognan, F. & Barber, C.（2020），《在临床前毒理学检测中引入虚拟对照组的概念》，《*ALTEX*—动物实验替代方案》**37**(3), 343–349。ter Riet, G., Korevaar, D. A., Leenaars, M., Sterk, P. J., Van Noorden, C. J. F., Bouter, L. M., Lutter, R., Elferink, R. P. O. & Hooft, L.（2012），《实验动物研究中的发表偏倚：对其程度、驱动因素、后果及潜在解决方案的调查》，*PLOS ONE* **7**(9), 1–5。**网址：***https://doi.org/10.1371/journal.pone.0043404*Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G. & Higgins, J. P.（2015），《研究间异质性的预测分布及其在贝叶斯荟萃分析中的简易应用方法》，《医学统计学》**34**(6), 984–998。van der Naald, M., Wenker, S., Doevendans, P. A., Wever, K. E. & Chamuleau, S. A.（2020），《临床前研究的发表率：呼吁预注册》，*BMJ Open Science* **4**(1), e100051。van Walraven, C.（2010），《个体患者荟萃分析——益处与挑战》，《临床流行病学杂志》**63**(3), 235–237。Vehtari, A., Gelman, A. & Gabry, J.（2017），《使用留一交叉验证和WAIC进行实用的贝叶斯模型评估》，《统计与计算》**27**(5), 1413–1432。Voelkl, B., Altman, N. S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N. A., Kas, M. J., Schielzeth, H., Van de Casteele, T. 等（2020），《生物变异背景下动物研究的可重复性》，《自然综述：神经科学》**21**(7), 384–393。Voelkl, B. & Würbel, H.（2016），《可重复性危机：我们是否忽视了反应规范？》，《药理学趋势》**37**(7), 509–510。Wagenmakers, E.-J., Lee, M. D., Rouder, J. N. & Morey, R. D.（2020），《预测无关性原则，或为何不应使用区间比较包含点零假设的模型》，载于《心理学统计学理论》，施普林格，第111–129页。Walley, R., Sherington, J., Rastrick, J., Detrait, E., Hanon, E. & Watt, G.（2016），《在重复临床前体内研究中使用贝叶斯分析，以更有效地利用动物》，《药物统计学》**15**(3), 277–285。Wandel, S., Neuenschwander, B., Röver, C. & Friede, T.（2017），《使用II期数据分析III期研究：在罕见病中的应用》，*Clinical Trials***14**(3), 277–285。Weber, S., Li, Y., Seaman, J. W., Kakizume, T. & Schmidli, H.（2021），《使用R贝叶斯证据综合工具应用荟萃分析预测先验》，《统计软件杂志》**100**(19), 1–32。Welch, B. L.（1938），《总体方差不等时两个均值之差的显著性》，《生物计量学》**29**(3/4), 350–362。Welch, B. L.（1947），《当涉及多个不同总体方差时对“Student问题”的推广》，《生物计量学》**34**(1-2), 28–35。Yang, H. & Novick, S.（2019），《药物开发中的*R*贝叶斯分析：概念、算法与案例研究》，CRC出版社。Zeileis, A.（2006），《三明治估计量的面向对象计算》，《统计软件杂志》**16**(9), 1–16。

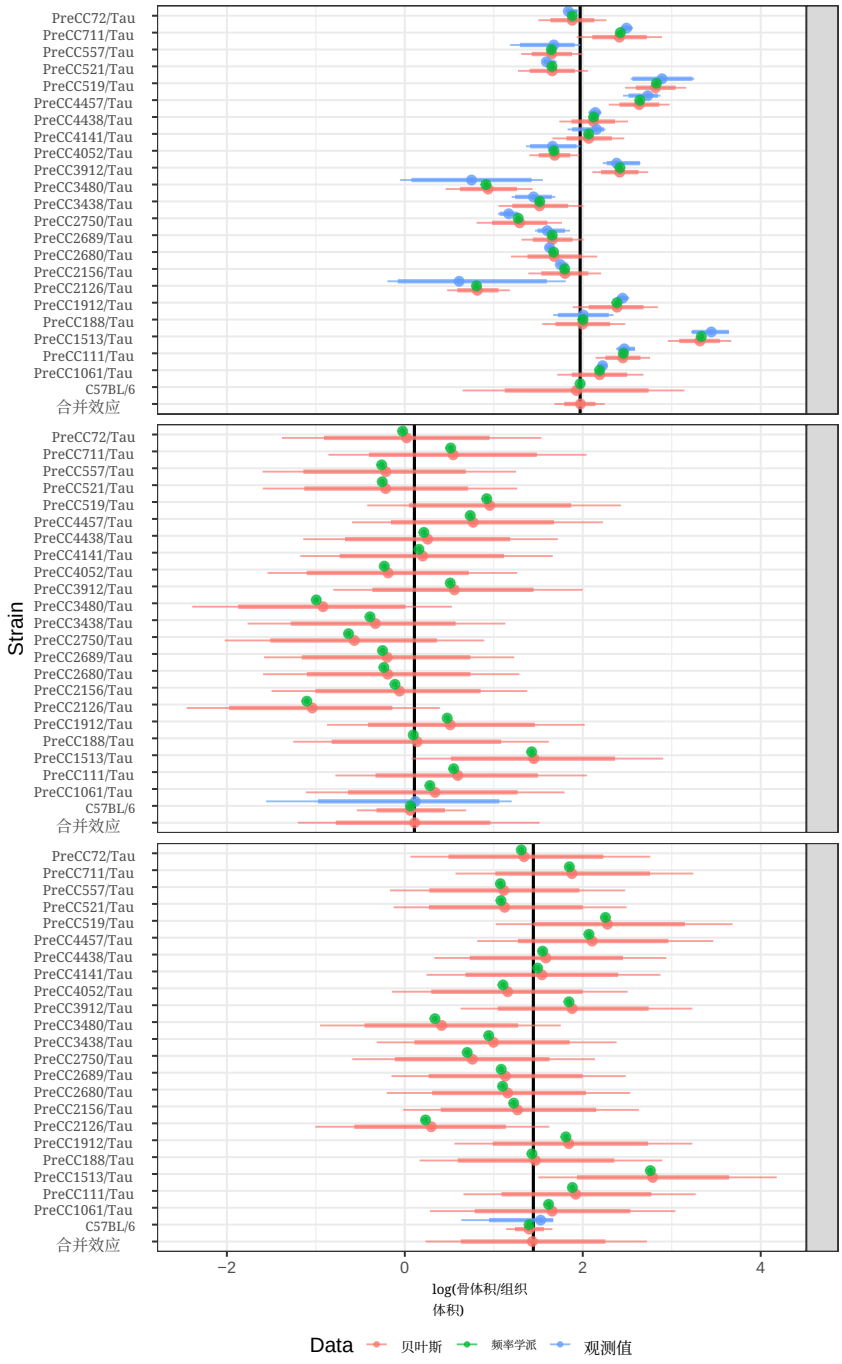


图2：按手术组（无手术、卵巢切除术（Ovx）、假手术）分层的历史动物品系特异性均值及其共同均值（合并效应）森林图。贝叶斯：后验MCMC抽样值的均值，即贝叶斯估计（点）。频率学派：REML估计。观测值：观测数据中的品系特异性均值。区间以 80%（粗线）和 85%（细线）分位数区间表示。

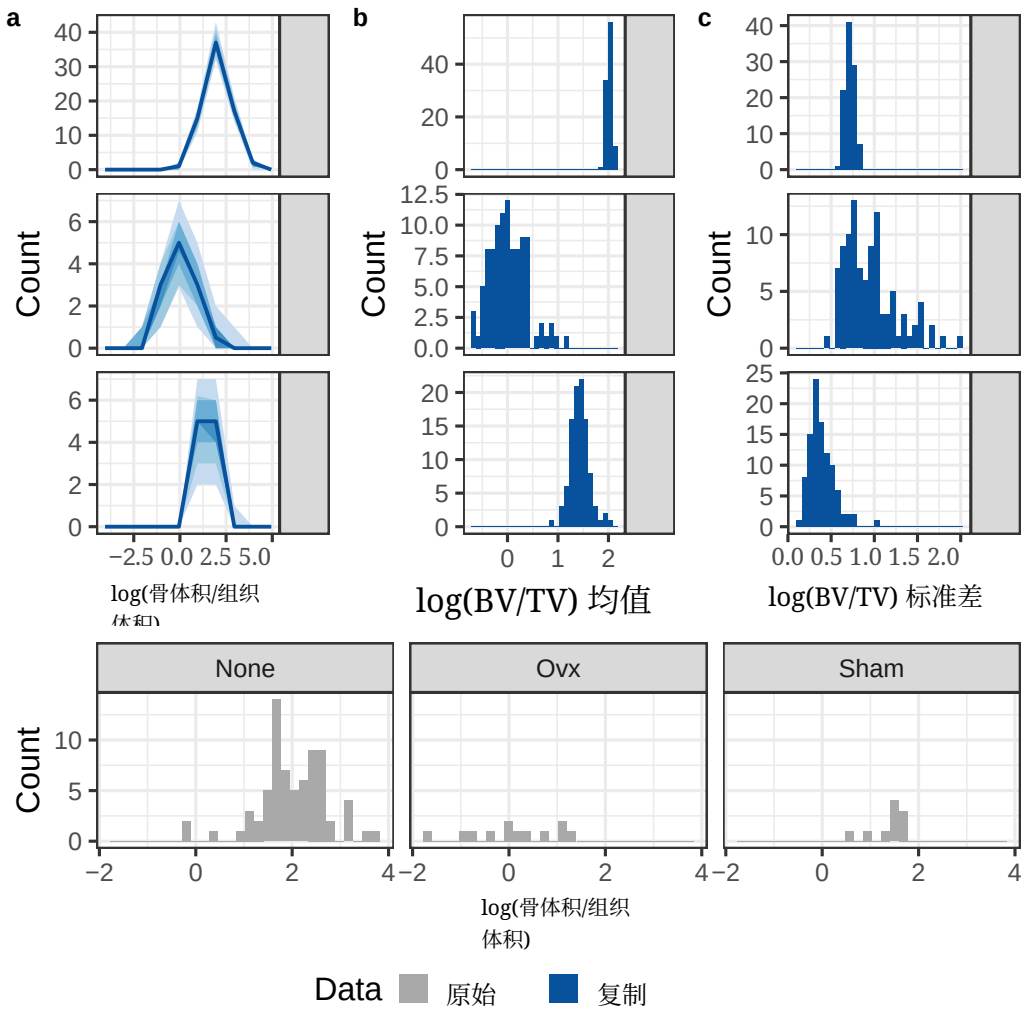
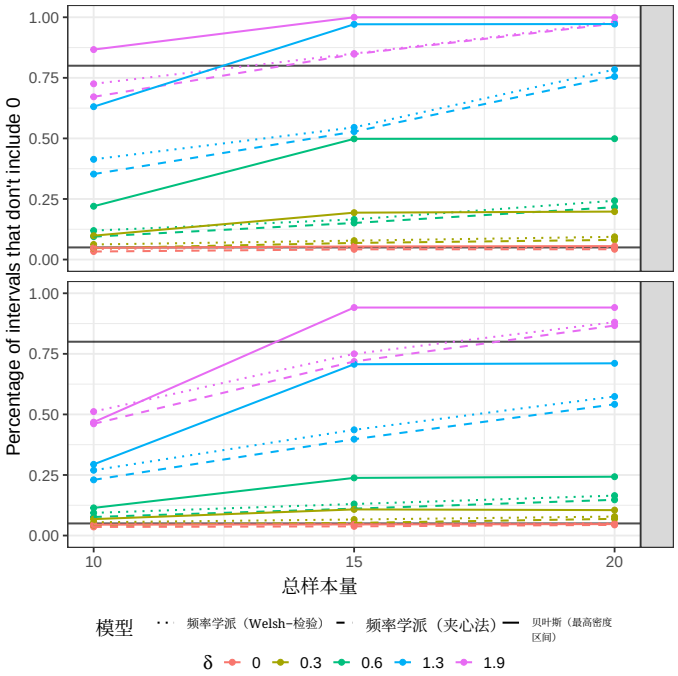
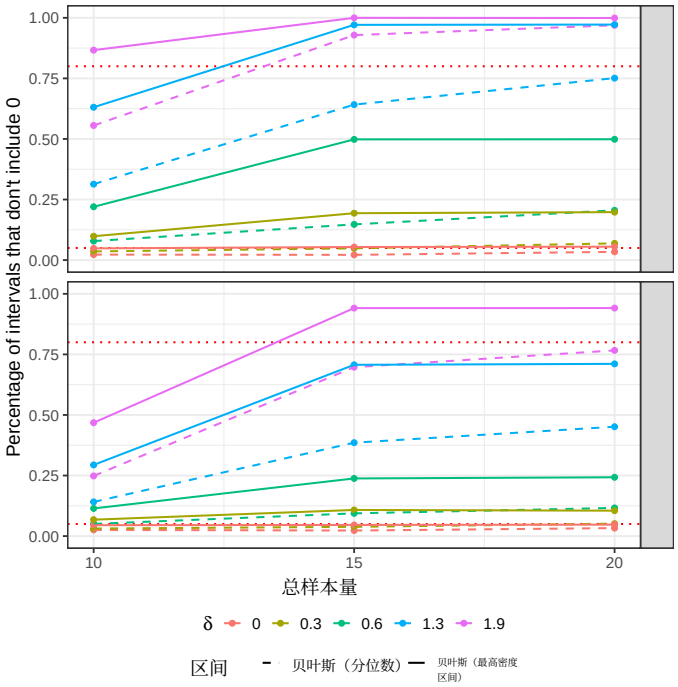


图3：改编自 Schad 等人（2021年）的图形化后验预测检验，展示了未接受手术的动物的相对骨体积（对数尺度）。灰色：原始数据。蓝色：来自马尔可夫链蒙特卡罗拟合的重复样本，基于1000个模拟数据集。a）每个模拟数据集计算所得直方图的分布。彩色区域按颜色强度递增的顺序，分别对应所有模拟数据集直方图频数的10–90、20–80、30–70和40–60百分位区间。区间中间的深色曲线表示所有模拟数据集的中位数分布。b）算术平均数的分布。c）标准差的分布。极端 log(骨体积/组织体积) 值 < -50 或 > 50 以 -50 和 50 表示。



(a)



(b)

图4：满足“成功”判定标准的模拟数据集所占比例：95%置信区间或可信区间不包含零值0，或Welch检验的p值小于0.05。分布根据10000个模拟数据集计算。(a)：点线：采用双侧Welch检验时，模拟数据集中p值< 0.05的比例。虚线：使用HC3夹心估计量的频率学派置信区间不包含0的模拟数据集比例。实线：贝叶斯最高密度区间（最高密度区间）不包含0的模拟数据集比例。(b)：比较贝叶斯分位数区间和最高密度区间。点线：I类错误率或检验效能为0.05和0.8时的常规界限。

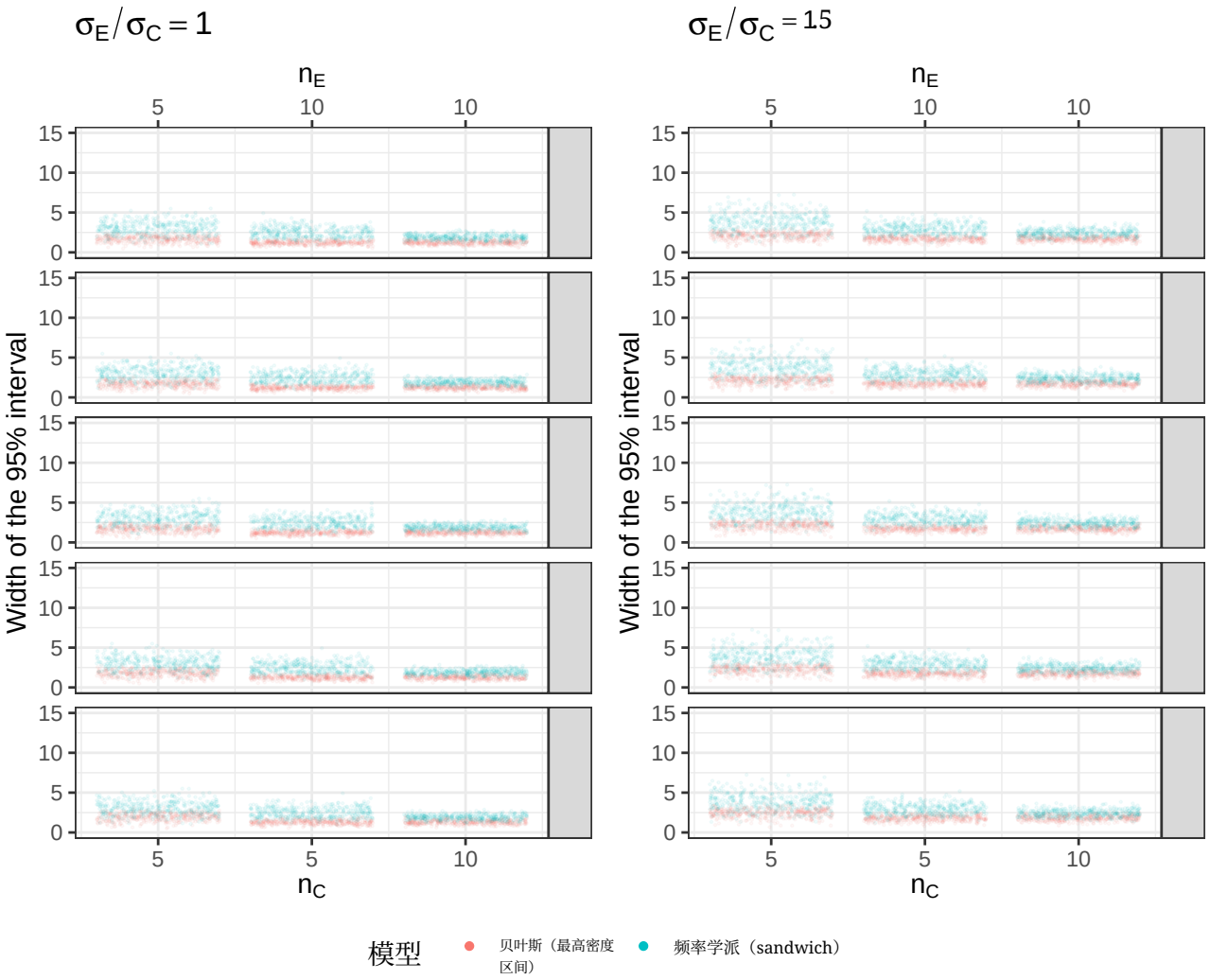


图5：.....的宽度 95% 贝叶斯最高密度区间（最高密度区间）和 95% 频率学派置信区间（采用 HC3 sandwich 估计量）针对不同设置下模拟数据集中的 500 个子样本。

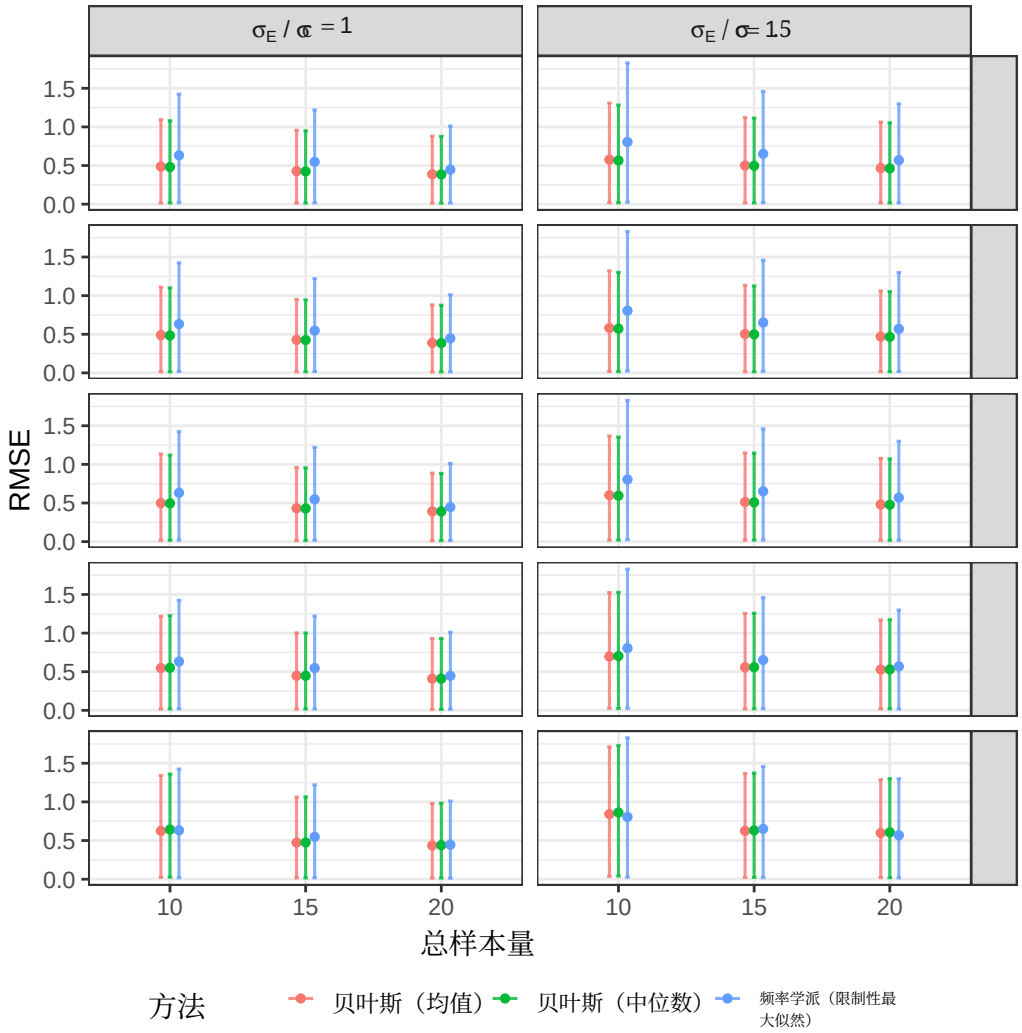


图 6：均方根误差（RMSE）计算为处理效应点估计与处理效应真实均值之差的平方平均值的平方根（$E(\delta)$）。RMSE 表示为所有模拟数据集的平均值，并附有 95%分位数区间。在贝叶斯模型中，δ 的点估计为 δ 后验 MCMC抽样值的算术平均数和中位数；在频率学派模型中，δ 的估计值则计算为处理组与对照组均值之差。

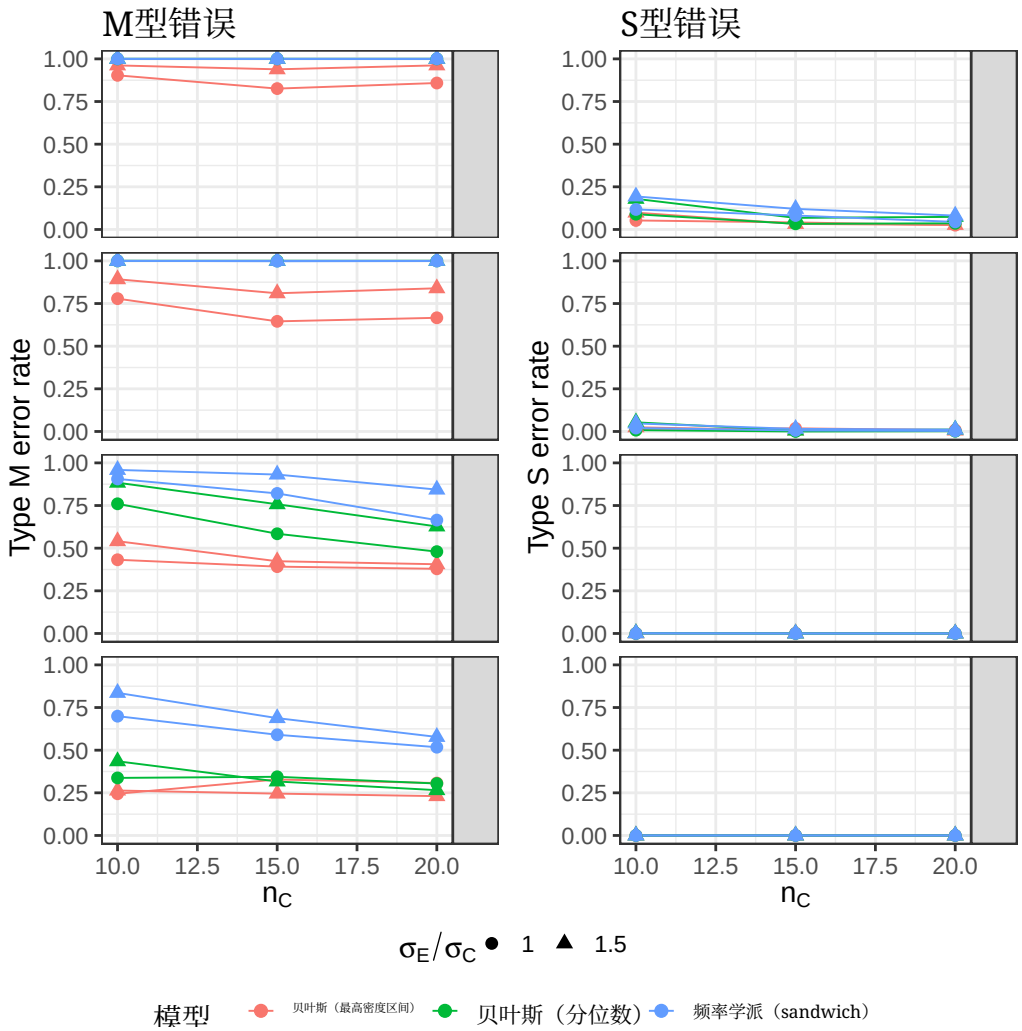


图7：M型（幅度）错误率：在置信区间/可信区间不包含0的模拟数据集中，效应估计值的绝对值大于真实处理效应 δ 的模拟数据集所占百分比。S型（符号）错误率：在置信区间/可信区间不包含0的模拟数据集中，效应估计值的符号与真实处理效应 δ 的符号不同的模拟数据集所占百分比。这些分布基于10000个模拟数据集计算得出。

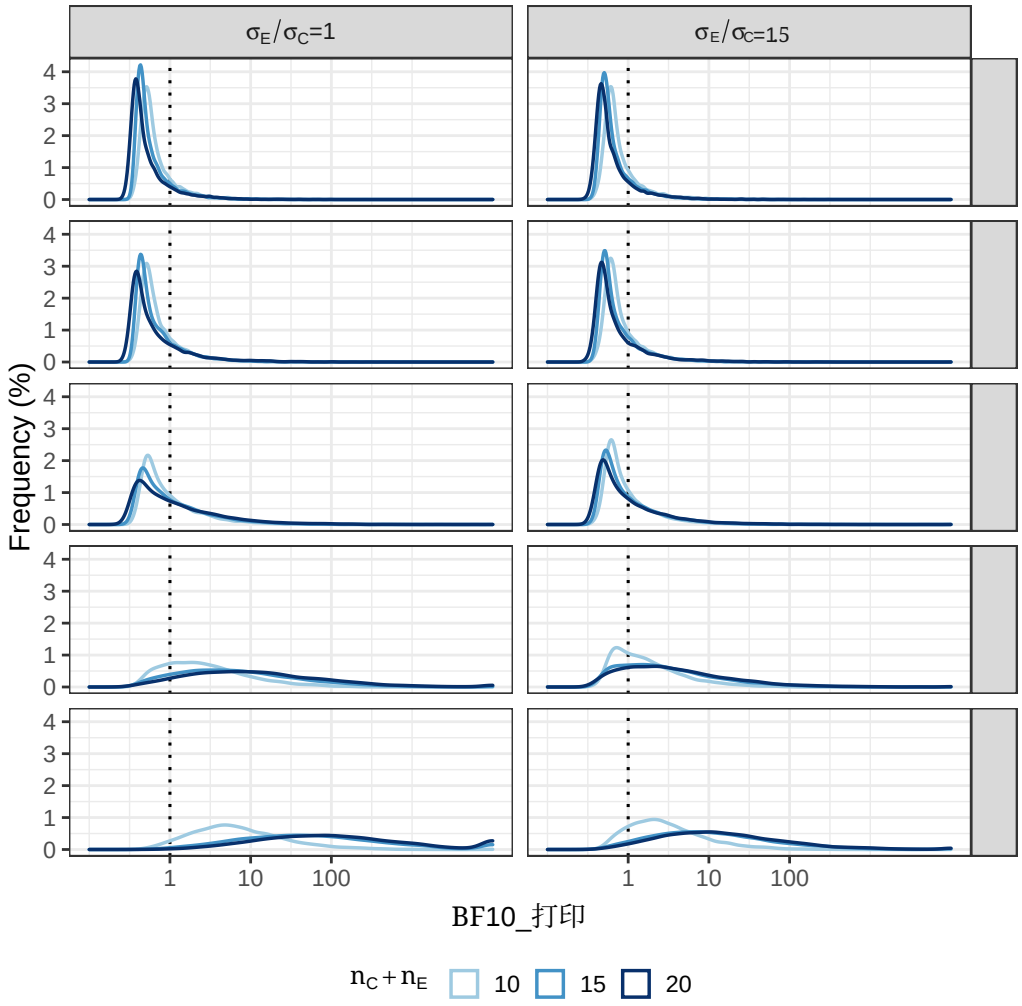


图8：估计的贝叶斯因子 BF_{10} 的分布，用于衡量支持处理效应 δ 不等于零的备择模型相对于处理效应等于零的零模型的证据。针对不同设计，基于数据中模拟的真实分布，在 10000 个模拟数据集上计算这些分布；每种设计均模拟 ˆ 10000 个数据集。为便于表示，$BF_{10} > 10000$ 的极端值以 10000 呈现。