ベイズ流メタアナリシスにおける
予測区間の評価と影響力解析

濱口　雄太

博士（学術）

総合研究大学院大学
複合科学研究科
統計科学専攻

令和４（２０２２）年度

概要

メタアナリシスとは，過去に行われた研究から得られたエビデンスを統合し，総合的な治療効果の評価を行うための方法である．近年では，ベイズ流の分析方法が，標準的な方法の一つとして普及しつつあり，医学研究の実践においても広く用いられている．本研究では，近年のベイズ流メタアナリシスの方法論において重要な問題とされている 2 つの研究課題について，技術評価・方法開発を行った．第一に，メタアナリシスの変量効果モデルにおいて，試験間の異質性を評価するための指標として，近年，急速に普及している予測区間の技術評価を行った．無情報事前分布を用いたもとでのベイズ流の予測区間は，頻度論的予測区間を近似するための有効な方法であると考えられていたが，現実的な条件下で，頻度論的な意味での正確な被覆確率を持つ保証はなく，また，それを示したエビデンスも，これまでに皆無である．本研究では，広範な実践的条件下でのシミュレーション実験を行い，その頻度論的な性能の詳細な分析を行った．第二に，外れ値の検出およびその影響力解析についての方法の開発を行った．ベイズ流のメタアナリシスでは，変量効果モデルを用いた階層ベイズモデルによって，試験間の異質性をモデル化し，総合的な治療効果の分析を行うことが一般的であるが，その中でも，特に極端なプロファイルを持つ試験が含まれる場合には，統合解析の結果に対して，深刻な影響を生じさせるリスクを有する．本研究では，Carlin-Louis 式の枠組みをもとに，ベイズ流メタアナリシスにおける，外れ値の検出と影響力解析の方法の開発を行った．いずれの研究においても，最新のシステマティックレビューの事例をもとに，その有用性の詳細な分析を行った．

目次

第 1 章 はじめに

医療や健康に関する科学的知見は，社会や行政・産業界から，我々の生活そのものにまで大きな影響を与えるものが多く，世界中で日々，膨大な研究成果が蓄積されている．例えば，米国国立生物工学情報センター（NCBI）が運営する文献データベース PubMed には，年間 100 万件以上の膨大な数の新たな文献が登録されている．システマティックレビュー（systematic review）とは，こうして蓄積・更新されていくエビデンスを網羅的に収集し，系統的な評価および統合解析を行い，最新・最善の医療情報を構築することを目的として行われるレビューのことをいう．そして，そのために用いられるエビデンス統合のための統計解析の方法のことをメタアナリシス（meta-analysis）という．

メタアナリシスにおいては，近年までの計算技術の飛躍的な発展と計算機性能の向上により，ベイズ流の方法が一つの標準的な方法として広く用いられている(Higgins et al., 2019; Spiegelhalter et al., 2004)．ベイズ流の解析手法の大きな利点は，マルコフ連鎖モンテカルロ法の活用により，柔軟な階層構造のモデルを用いた解析が可能であるという点と，事前情報を積極的に解析に取り込むことができるという点である．近年では，ソフトウェアの開発・整備も進められており，R 等の統計ソフトウェアを用いることによって，簡単なコマンドで，非統計家が高度なベイズ流の解析を実行することもできるようになっている(Röver, 2020)．一方で，従来の主たる方法であった頻度論に基づく方法に比べて，ベイズ流の方法には，実践上，重要な手法に未整備・未確立のものがあったり，十分な技術評価が行われないまま，実践で広く用いられている方法もある．システマティックレビューから得られるエビデンスは，医療技術評価や診療ガイドラインの作成，医療政策の策定等にも広く用いられるため，これらの方法論的課題を解決することは，極めて重要な課題となる．

本研究では，近年のメタアナリシスの方法論において重要な問題とされている 2 つの課題についての技術評価・方法開発を行った．第一に，メタアナリシスの変量効果モデルにおいて，試験間の異質性を評価するための指標として，近年，急速に普及した予測区間（prediction interval）の技術評価を行った．予測区間は，Higgins et al. (2009)により，将来の研究における真の治療効果が分布する範囲を異質性の評価に用いるものとして提案されたが，最近の研究によって，Higgins et al. (2009)による予測区間の構成方式では，統計的な誤差が過小評価されることが明らかにされた(Nagashima et al., 2019; Partlett and Riley, 2017). 一方，無情報事前分布を用いたもとでのベイズ流の方法が，その代替的な方法となると考えられ(Higgins et al., 2009)，多くの研究で，ベイズ流の予測区間が提示されているが，そうして与えられる予測区間が，頻度論的な意味での正確な被覆確率を持つものとなる保証はなく，また，それを明確に分析したエビデンスも，これまでに皆無である．本研究では，広範な実践的条件下でのシミュレーション実験を行い，その頻度論的な性能の詳細な分析を行った．第二に，外れ値の検出およびその影響力解析についての方法の開発を行った．ほとんどのシステマティックレビューでは，対象となる試験間で，さまざまな要因が異質であり，試験間の治療効果の大きさには異質性が存在する(Higgins et al., 2019). ベイズ流のメタアナリシスでは，変量効果モデルを用いた階層ベイズモデルによって，この異質性をモデル化し，総合的な治療効果の分析を行うことになるが，この中でも，特に極端なプロファイルを持つ試験が含まれる場合には，その試験は，統合解析の結果に対して，深刻な影響を生じさせるリスクを有する．頻度論の枠組みにおいては，そのような外れ値（outlier）となる試験の検出と影響力解析を行うための方法が開発されてきたが(Viechtbauer and Cheung, 2010)，ベイズ流の枠組みにおいては，そのような方法は，これまで議論されていなかった．本研究では，ベイズ流メタアナリシスの枠組みにおい

て，Carlin and Louis (2009)の式の外れ値の検出と影響力解析の方法についての開発を行った．

本論文の構成は，以下の通りである．まず，第 2 章で，本研究で用いるメタアナリシスのモデルとベイズ流の分析手法について概説する．続いて，第 3 章において，ベイズ流メタアナリシスの予測区間の技術評価研究について述べる．そして，第 4 章において，Carlin-Louis 式の外れ値の検出と影響力解析の方法の開発研究について述べる．最後に，第 5 章において，本研究の結論を述べる．

第 2 章 メタアナリシス

2.1 変量効果モデル

メタアナリシスのモデルとして，各試験の治療効果の違いである異質性が存在しないと仮定する固定効果モデルと，異質性が存在すると仮定する変量効果モデルが提案されている(DerSimonian and Laird, 1986; Whitehead and Whitehead, 1991)．異質性が存在しないとする仮定はメタアナリシスの事例のデータから検証できず，仮定が誤っている場合に平均治療効果の信頼区間が過小評価されることで誤った結論を導く可能性がある．そのため，異質性が存在すると仮定する変量効果モデルが一般的に使用されており，本論文においても使用する．変量効果モデルは以下のように定義される．

$$Y_k \sim N(\theta_k, \sigma_k^2)$$

$$\theta_k \sim N(\mu, \tau^2)$$

ここで，統合試験数をK, 試験kの治療効果の確率変数を$Y_k (k = 1,2, \dots, K)$とする．θ_kは試験kの治療効果，μは平均治療効果，σ_k^2は試験内分散，およびτ^2は試験間分散を表す．一般的に用いられる治療効果の指標は，平均差，標準化平均差，リスク差，リスク比，ハザード比，およびオッズ比である(Higgins et al., 2019; Whitehead, 2002)．比の指標は，一般的な臨床試験の規模のサンプルサイズでは漸近正規性の近似が不十分であるため，正規近似の精度を上げるために一般的に対数変換して用いられる．なお，$\tau^2 = 0$である変量効果モデルは，固定効果モデルと一致する．

変量効果モデルにより，事例のデータから平均治療効果および異質性の指標である試験間分散を推定する．頻度論的な手法では，平均治療効果および試験間分散をモーメント法や制限付き最尤推定法等を使用して推定する(Higgins et al., 2019)．平均治療効果の推定量は以下のように計算される．

$$\hat{\mu} = \frac{\sum_{k=1}^{K} \widehat{w}_k Y_k}{\sum_{k=1}^{K} \widehat{w}_k}$$

ここで，$\widehat{w}_k = (\sigma_k^2 + \hat{\tau}^2)^{-1}$とする．

また，試験間分散の推定量に関して，モーメント法を使用したDerSimonian-Laird推定量(DerSimonian and Laird, 1986)は，計算が容易で多くの統計ソフトウェアで計算することができるため，メタアナリシスの実践で使用されている．

$$\hat{\tau}_{DL}^2 = \max[0, \hat{\tau}_{UDL}^2], \hat{\tau}_{UDL}^2 = \frac{Q - (K-1)}{S_1 + S_2/S_1}$$

$$Q = \sum_{k=1}^{K} \frac{(Y_k - \bar{Y})^2}{\sigma_k^2}, \bar{Y} = \frac{\sum_{k=1}^{K} \sigma_k^{-2} Y_k}{\sum_{k=1}^{K} \sigma_k^{-2}}$$

$$S_r = \sum_{k=1}^{K} (\sigma_k^{-2})^r, r = 1,2$$

平均治療効果や試験間分散の推定量を用いて，平均治療効果の$100(1-\alpha)\%$信頼区間は一般的に以下のように定義される．

$$\left[\hat{\mu} - t_{K-1}^{\alpha}\sqrt{\widehat{Var}[\hat{\mu}]}, \hat{\mu} + t_{K-1}^{\alpha}\sqrt{\widehat{Var}[\hat{\mu}]}\right]$$

ここで，t_{K-1}^{α}は自由度$K-1$のt分布における$100(1-\alpha/2)\%$分位点とする．この信頼区間は統合する試験数が無限大とする大標本近似を使用しているが，医学研究の大半のメタアナリシスの試験数は 20 以下であり，大標本近似が成立しないものと考えられる．実践的な設定のもとでのシミュレーション実験による信頼区間の評価では，被覆確率が名目水準を下回ることが報告されている(Veroniki et al., 2019)．この大標本近似による問題は第 3 章で示す予測区間においても同様に起こる．

2.2 ベイズ流メタアナリシス

DerSimonian-Laird の方法等の従来の頻度論的な手法では変量効果モデルの試験間分散の点推定量を使用しており，その不確実性を考慮することができない．Higgins らは信頼区間や予測区間の推定に試験間分散の不確実性を考慮できるベイズ流の手法を推奨しており，ベイズ流メタアナリシスが実践で普及しつつある(Higgins et al., 2009).

ベイズ流メタアナリシスにおいても頻度論的な手法と同様に変量効果モデルを使用する．以下のベイズの定理に従い，パラメータの事前分布と事例のデータにより事後分布を推定する．

$$p(\boldsymbol{\xi}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\xi})p(\boldsymbol{\xi}) = \prod_{k=1}^{K} p(Y_k|\theta_k, \sigma_k^2)p(\theta_k|\mu, \tau^2)p(\mu, \tau^2)$$

ここで，$\boldsymbol{\xi}$は全てのパラメータ，$\boldsymbol{y}$は全ての試験のデータとする．平均治療効果と試験間分散を客観的に推定するためには，ベイズ推定における事前分布として無情報事前分布が使用される．一般的に，平均治療効果では分散の大きい正規分布等が使用され，試験間分散では一様分布や一様であると近似できる逆ガンマ分布が使用されている(Lambert et al., 2005)．事後分布の推定には，R パッケージの bayesmeta (Röver, 2020)や OpenBUGS 等を用いたマルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo methods; MCMC）(Lunn et al., 2009)を利用することができる．

第 3 章 ベイズ流予測区間の頻度論的性能評価

3.1 背景と目的

メタアナリシスにおける変量効果モデルは，各試験の治療効果の違いである異質性を考慮できるため，標準的な分析手法の一つとなっている(DerSimonian and Laird, 1986; Whitehead and Whitehead, 1991)．異質性を定量する指標として，全分散に対する試験間分散の割合を推定するI^2統計量が提案されているが(Higgins and Thompson, 2002)，平均治療効果とI^2統計量が別々に提示されることで結果の解釈が困難であることが問題とされている(Riley et al., 2011)．そこで，平均治療効果と異質性を統合した指標として予測区間が提案された(Higgins et al., 2009)．予測区間は，将来に実施すると想定される試験における真の治療効果が一定の確率により含まれる区間として定義され，臨床的な解釈が容易であり，平均治療効果と異質性の不確実性を同時に評価できることから，近年のメタアナリシスの方法論研究においてその有用性が議論されている(IntHout et al., 2016; Veroniki et al., 2019)．Higgins et al. (2009)によって提案されたモーメント法を使用する Higgins-Thompson-Spiegelhalter（HTS）法は予測区間を計算する方法として最も広く使用されている．この HTS 法は計算が簡便で実践で使用しやすいが，メタアナリシスで統合する試験数が小さい場合，予測区間の被覆確率が過小になることが報告されている(Partlett and Riley, 2017)．その理由の最も重要なものとして，試験数が十分に大きいという大標本近似を使用している点が挙げられる．医学研究におけるメタアナリシスでは，試験数は 20 以下であることが大半であり(Kontopantelis et al., 2013)，この大標本近似は，一般的に成立しないことが想定される(Brockwell and Gordon, 2001; Noma, 2011; Veroniki et al., 2019)．HTS 法の予測区間の被覆確率が過小になることで治療効果が誤って解釈されることは，重大な問題であると考えられる．

この問題に対する有効な代替的な手法として，ベイズ流の手法が挙げられる(Carlin and Louis, 2009; Gelman et al., 2013)．HTS 法を提案した Higgins et al. (2009) もベイズ流の手法をメタアナリシスにおいて予測区間を構築する有効なアプローチとして提案している．ベイズ流の手法は，その公理に基づき，純粋に主観的な手法として使用される場合，その推測・予測は正確なものとなるが(Bodnar et al., 2017; Röver, 2020)，無情報事前分布を用いて，頻度論的方法の近似手法として使用されることも多い．実際，サンプルサイズが十分に大きい条件下では，無情報事前分布を用いたベイズ流の手法から得られる推測・予測の結果は，頻度論的な分析の良い近似を与えることが示されている(Carlin and Louis, 2009; Gelman et al., 2013; Spiegelhalter et al., 2004)．そのため，メタアナリシスにおいても，実質的には頻度論的な推測・予測の近似手法として，ベイズ流の手法が用いられることが多いように考えられる．しかしながら，頻度論的な予測区間は，将来の観測値がある客観的な確率を持って含まれる区間として厳密に定義されており，ベイズ流の主観確率に基づく予測とは，根本的にその概念が異なる．また，ベイズ流の信用区間を評価したシミュレーション実験では無情報事前分布を用いたベイズ流の推測は必ずしも頻度論的な推測の近似とならないことが示されており(Agresti and Min, 2005)，事前分布の種類により近似性能が大幅に変わることも示唆されている(Lambert et al., 2005)．無情報事前分布を使用したベイズ流の予測区間が HTS 法の代替手段として提案されているが，これまで実践的なメタアナリシスの条件下におけるベイズ流の予測区間の性能をシミュレーション実験により評価した明確なエビデンスは存在しない状況である．

本章では，ベイズ流の予測区間の特性に関する定量的なエビデンスによりこれらの手法の実践的な使用を推奨するかどうかを提示するために，11 種類の多様な無情報事前分布を使用したベイズ流の予測区間の頻度論的な性能を広範な

シミュレーション実験により評価した．さらに，近年出版された主要な医学雑誌におけるメタアナリシスの 2 事例に適用した．

3.2 既存の頻度論的手法

メタアナリシスにおける$100(1-\alpha)\%$予測区間は，同様の母集団を想定して将来 1 つの試験を実施した際の治療効果θ_{new}が$100(1-\alpha)\%$の確率で含まれ得る範囲として一般的に定義される(Higgins et al., 2009; Riley et al., 2011)．Higgins et al. (2009)は予測区間を簡便に計算する頻度論的な手法として HTS 法を提案した．

$$\left[\hat{\mu}-t_{K-2}^{\alpha}\sqrt{\hat{\tau}_{DL}^2+\widehat{Var}[\hat{\mu}]},\hat{\mu}+t_{K-2}^{\alpha}\sqrt{\hat{\tau}_{DL}^2+\widehat{Var}[\hat{\mu}]}\right]$$

ここで，$\hat{\tau}_{DL}^2$は試験間分散τ^2のモーメント推定量である DerSimonian-Laird 推定量(DerSimonian and Laird, 1986)，$\widehat{Var}[\hat{\mu}]=1/(\sum_{k=1}^{K}(\sigma_k^2+\hat{\tau}_{DL}^2)^{-1})$は平均治療効果の推定量の分散推定量，$t_{K-2}^{\alpha}$は自由度$K-2$の$t$分布における$100(1-\alpha/2)\%$分位点とする．

HTS 法は 2 種類の近似を使用している．

$$(\hat{\mu}-\mu)/\sqrt{\widehat{Var}[\hat{\mu}]}\sim N(0,1)$$

$$(K-2)(\hat{\tau}_{DL}^2+\widehat{Var}[\hat{\mu}])/(\tau^2+\widehat{Var}[\hat{\mu}])\sim\chi^2(K-2)$$

これらの近似はメタアナリシスにおいて試験数が十分に多い大標本である場合に成立する．しかし，医学研究におけるメタアナリシスでは，試験数は 20 以下であることが大半であり(Kontopantelis et al., 2013)，Partlett and Riley (2017)や Nagashima et al. (2019)は HTS 法の予測区間の被覆確率が医学研究のメタアナ

リシスの一般的な条件下で名目水準より低くなり，過小評価されることをシミュレーション実験により示している．

3.3 ベイズ流の手法

3.3.1 階層ベイズモデル

マルコフ連鎖モンテカルロ法（MCMC）を用いるベイズ流の手法は統計学における予測方法の一つとして確立されており，統計的予測問題における代表的な手法である．Higgins et al. (2009)は，メタアナリシスにおける予測問題を解決する手段としてベイズ流の手法を提示している．Higgins et al. (2009)が提案したベイズ流の手法では，新たに行う試験の治療効果の予測分布は，観測された試験結果が所与のもとで，平均がμ，分散がτ^2である正規分布に従うθ_{new}をMCMCによりサンプリングすることにより計算される(Higgins et al., 2009; Smith et al., 1995)．新たな試験の95%予測区間はθ_{new}の事後予測分布の2.5%分位点から97.5%分位点までの区間として得られる．

変量効果モデルにおける階層ベイズモデルは一般的には変量効果分布のパラメータμおよびτ^2の事前分布を仮定して構築される．純粋なベイズ流の予測法はこの枠組みで実行されるが，本章ではこのベイズ流の枠組みについて無情報事前分布を使うことで近似的な頻度論的な予測に使用する．

3.3.2 事前分布

メタアナリシスのベイズモデルにおいては，様々な無情報事前分布が考案されている．Lambert et al. (2005)はシミュレーション実験によりμやτ^2の頻度論的な推定性能を評価し，τ^2の事前分布に大きく依存すると結論付けた．そこで，

本章では非正則（improper）な事前分布を含む 11 種類の無情報事前分布の予測性能に対する影響を評価することとした.

本章では，μとτの同時分布をそれぞれの周辺分布の積により得られると仮定した．平均治療効果μの事前分布を以下の通り，仮定した.

$$\mu \sim N(0,10000)$$

試験間分散τ^2の事前分布として，以下の 11 種類の分布を使用した.

τに対する一様な事前分布 (Uniform)

$$p(\tau) \propto 1$$

この分布はτに対して一様で最も直感的な，平坦な事前分布である(Röver, 2020).

$\sqrt{\tau}$に対する一様な事前分布 (Sqrt)

$$p(\tau) \propto \frac{1}{\sqrt{\tau}}$$

この分布は$\sqrt{\tau}$に対しての一様な事前分布である．τの変数変換に対する不変性があるため，合理的な無情報事前分布である(Jaynes, 1968, 2003). また，$p(\tau) \propto \tau^a\ (-\infty < a < \infty)$における$a = -0.50$である分布であり，密度が単調に減少する．なお，$a = 0$である分布は Uniform と一致する.

Jeffreys 事前分布 (Jeffreys)

$$p(\tau) \propto \sqrt{\sum_{k=1}^{K} \left(\frac{\tau}{\sigma_k^2 + \tau^2} \right)^2}$$

この分布は尤度のフィッシャー情報量の平方根に比例する(Jeffreys, 1946). 形式的にはμとτ^2の同時分布として与えられるが，階層ベイズモデルではフィッシャー情報行列の対角成分以外が 0 であることによりμとτ^2が独立であると仮定している．この分布は，Tibshirani の無情報事前分布や Berger–Bernard の参照事前分布と一致する(Bodnar et al., 2017; Tibshirani, 1989).

Berger–Deely 事前分布 (Berger–Deely)

$$p(\tau) \propto \prod_{k=1}^{K} \left(\frac{\tau}{\sigma_k^2 + \tau^2} \right)^{1/K}$$

この分布は Jeffreys と同様に試験内分散を使用している(Berger and Deely, 1988). Jeffreys が相加平均を使用しているのに対して Berger-Deely は相乗平均を使用しており，試験内分散が全て等しい場合，Jeffreys と等しくなる．

定型的な事前分布 (Conventional)

$$p(\tau) \propto \prod_{k=1}^{K} \left(\frac{\tau}{(\sigma_k^2 + \tau^2)^{3/2}} \right)^{1/K}$$

この分布も Berger-Deely と同様に Jeffreys から派生した分布である(Berger and Deely, 1988). 検定やモデル選択の目的で使用される(Berger and Deely, 1988; Berger and Pericchi, 2001).

DuMouchel 事前分布 (DuMouchel)

$$p(\tau) = \frac{s_0}{(s_0 + \tau)^2}, \quad s_0^2 = \frac{K}{\sum_{k=1}^{K} \sigma_k^{-2}}$$

この分布は DuMouchel and Normand (2000)により提案され，試験内分散の調和平均s_0^2を使用し，最頻値が 0，中央値がs_0である対数ロジスティック分布である．

縮小事前分布 (Shrinkage)

$$p(\tau) = \frac{2s_0^2\tau}{(s_0^2+\tau^2)^2}, \quad s_0^2 = \frac{K}{\sum_{k=1}^{K}\sigma_k^{-2}}$$

この分布は平均パラメータのベイズ推定量の縮小因子である$S_0(\tau) = s_0^2/(s_0^2+\tau^2)$に対して一様である(Daniels, 1999; Spiegelhalter et al., 2004). DuMouchel と同様に試験内分散の調和平均s_0^2を使用している．

I^2統計量に対する一様な事前分布 (I2)

$$p(\tau) = \frac{2\hat{\sigma}^2\tau}{(\hat{\sigma}^2+\tau^2)^2}, \quad \hat{\sigma}^2 = \frac{(K-1)\sum_{k=1}^{K}\sigma_k^{-2}}{(\sum_{k=1}^{K}\sigma_k^{-2})^2 - \sum_{k=1}^{K}\sigma_k^{-4}}$$

この分布は Shrinkage における試験内分散の調和平均s_0^2を$\hat{\sigma}^2$に置き換えた，Higgins のI^2統計量に一様な分布である(Higgins and Thompson, 2002).

正則な事前分布 (Proper 1~3)

正則（proper）な無情報事前分布として$\tau \sim Uniform(0,10)$ (Proper 1), $1/\tau^2 \sim Gamma(0.001,0.001)$ (Proper 2), $1/\tau^2 \sim Gamma(0.01,0.01)$ (Proper 3)を使用する．Proper 1 は区間が制限された完全に一様な分布である．Proper 2 は試験間分散として最も一般的に使用される準共役事前分布である(Lambert et al., 2005). 大半の区間において近似的に一様であるが，0 近傍で密度が急上昇する．Proper 3 は Proper 2 と比較してやや情報を有する分布である．Proper 2 およ

び Proper 3 を使用した予測区間の頻度論的な性能を比較することでハイパーパラメータの変化による感度を評価することができると考えた.

3.4 シミュレーション実験

3.4.1 実験方法

ベイズ流の予測区間の頻度論的な性能を評価したエビデンスを与えるためにシミュレーション実験を実施した．前項に示した 11 種類の無情報事前分布および比較対照として Higgins et al. (2009)の予測区間（HTS）を評価した．将来の試験における 95%予測区間はθ_{new}の事後予測分布の 2.5%分位点から 97.5%分位点までの区間として得られる．計算には，事前分布として Proper 2 および Proper 3 を使用する場合，OpenBUGS を使用し(Lunn et al., 2009)，その他の事前分布を使用する場合，R パッケージの bayesmeta を使用した(Röver, 2020).

シミュレーションのデータはオッズ比を評価する医学研究における典型的なメタアナリシスを考慮した Brockwell and Gordon (2001, 2007)の設定を参考にして発生させた．予測区間の被覆や精度を評価するための一般性を失わないようにするため，平均治療効果μを 0 とした．試験内分散σ_k^2は自由度 1 のχ^2分布に従う乱数を 0.25 倍し，[0.009, 0.6]の範囲で打ち切った．このσ_k^2の分布の平均，中央値，95%信頼区間は，それぞれ，0.17，0.12，[0.01, 0.55]であった．なお，試験内分散はそれぞれのシミュレーションで別々に生成させた．試験数Kと試験間分散τ^2は 2 種類のパターンに変化させた．(1) 試験数を 7 および 15 に固定し，試験間分散を 0.01, 0.02,…, 0.20 と変化させた場合，(2) 試験数を 4, 5, 6,…, 20 と変化させ，試験間分散を 0.10 および 0.20 に固定させた場合である．12 種類の異なる予測区間で同じシミュレーションの試験データを使用した．それぞれのシナリオでシミュレーションを 10,000 回繰り返した．シミュレーション毎

に新たに行う試験の治療効果をθ_{new}~$N(\mu, \tau^2)$に従ってランダムに生成し，それらが予測区間に経験的に被覆される割合である被覆確率，および予測区間の期待値である期待区間幅を算出した．被覆確率は名目水準である 95%に一致することが望ましいとされる．

3.4.2 シナリオ 1 ($K = 7$)

試験数を 7 に固定したシナリオの結果を図 3.1 に示す．τ^2が 0.01 と小さい場合，ベイズ流の予測区間はいずれも被覆確率が高かった．一方，HTS 法による予測区間の被覆確率はτ^2が 0.01 の場合，名目水準の 95%付近に位置しているが，τ^2が上昇するに従い被覆確率が低下した．この HTS 法の傾向は過去のシミュレーション実験による報告と一致している(Nagashima et al., 2019; Partlett and Riley, 2017)．ベイズ流の予測区間はいずれもτ^2が上昇するに従い被覆確率が減少した．τ^2が 0.20 の場合，Sqrt，DuMouchel，Shrinkage，I2，Proper 2 は被覆確率が名目水準の 95%を下回り，Conventional，Proper 3 は被覆確率が名目水準の 95%付近に位置した．期待区間幅に関して，いずれの予測区間もτ^2が上昇するに従い増加した．

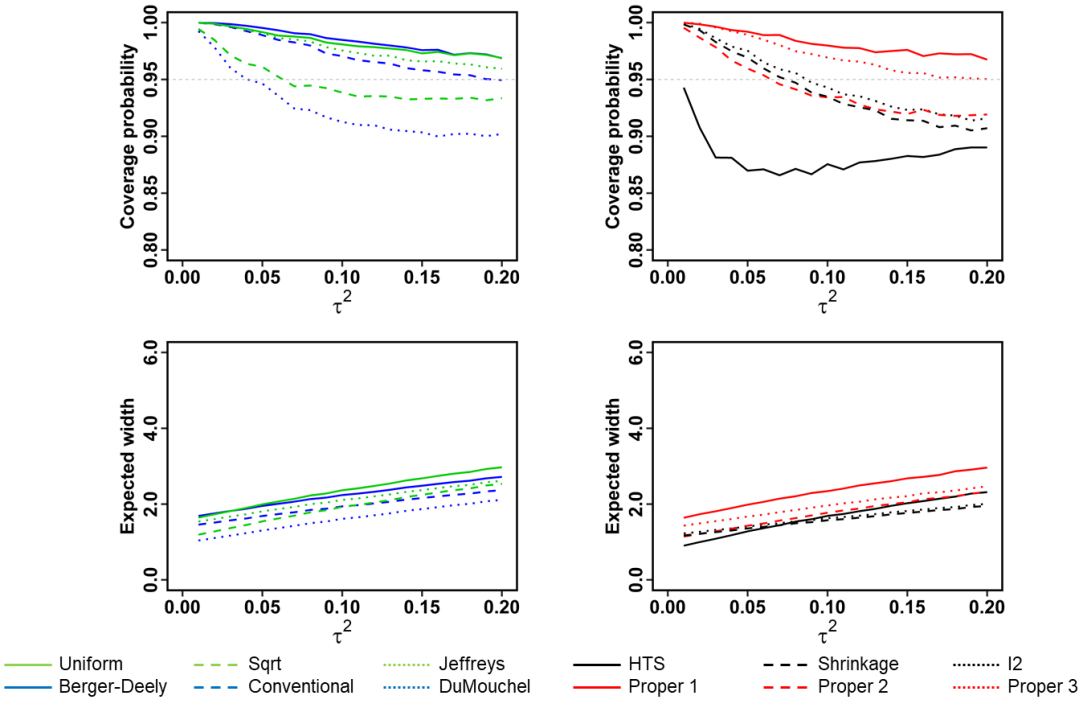

図 3.1　試験数を 7 に固定したシミュレーション実験における被覆確率および期待区間幅

3.4.3 シナリオ 2 ($K = 15$)

試験数を 15 に固定したシナリオの結果を図 3.2 に示す．τ^2が 0.01 と小さい場合，シナリオ 1 と同様にベイズ流の予測区間はいずれも被覆確率が 95%より高かった．一方，HTS 法による予測区間の被覆確率はτ^2が 0.01 の場合，名目水準の 95%付近を大幅に下回った．τ^2が 0.10 以下の場合，Uniform，Jeffreys，Berger-Deely，Conventional，Proper 1 は被覆確率が 95%より高いが，τ^2が 0.10 以上の場合，被覆確率が 95%付近に位置していた．Proper 2 および Proper 3 はいずれも逆ガンマ分布に従う事前分布であるが，被覆確率は大きく異なった．これらの結果から，いずれの予測区間が一般的に正確かを明示することは困難だが，Uniform，Jeffreys，Berger-Deely，Conventional，Proper 1 はτ^2が 0.10 以上の場合，正確な被覆確率を示すと考えられる．

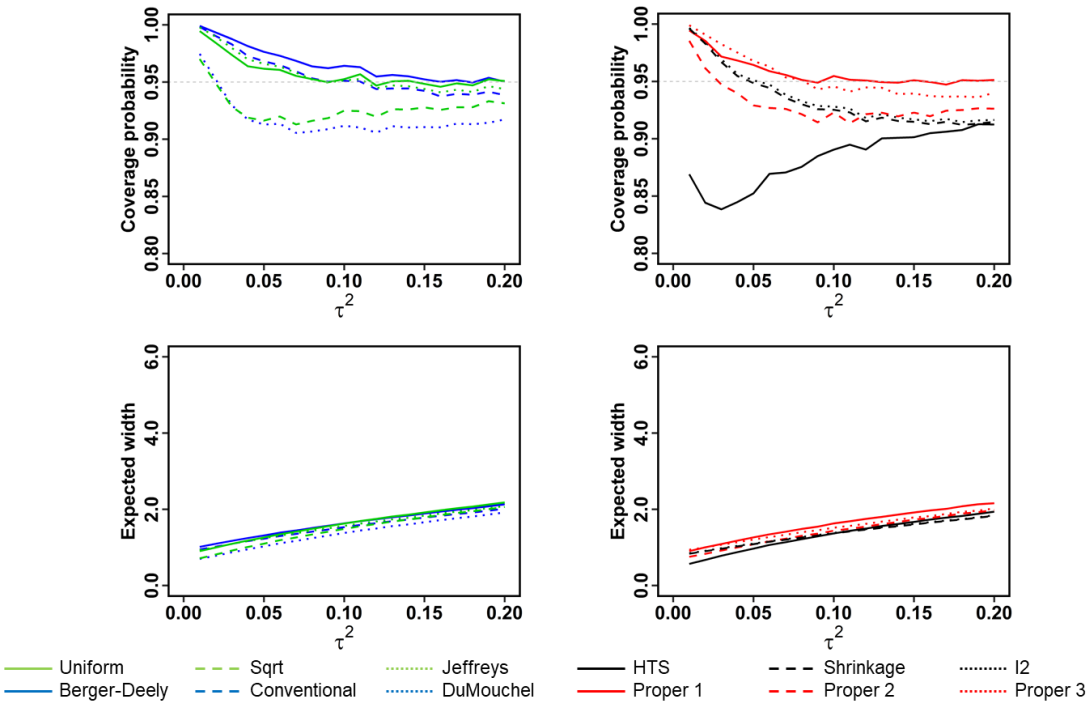

図 3.2　試験数を 15 に固定したシミュレーション実験における被覆確率および期待区間幅

3.4.4 シナリオ 3 ($\tau^2 = 0.10$)

試験間分散を 0.10 に固定したシナリオの結果を図 3.3 に示す．試験数が 4 あるいは 5 と少ない場合，DuMouchel，Shrinkage，I2，Proper 2，HTS が最も正確な被覆確率を示した．試験数が増加するに従い，これらの予測区間の被覆確率はいずれも減少し，特に DuMouchel は被覆確率が 90%まで減少した．Sqrt は試験数が 6 以下と少ない場合，被覆確率が 95%付近に位置するが，試験数が 7 以上の場合，95%を大幅に下回った．ベイズ流の予測区間の Uniform，Jeffreys，Berger–Deely，Conventional，Proper 1，Proper 3 は試験数が 11 以下の場合，被覆確率が 95%を上回るが，試験数が 12 以上の場合，概ね適切な被覆確率を示した．

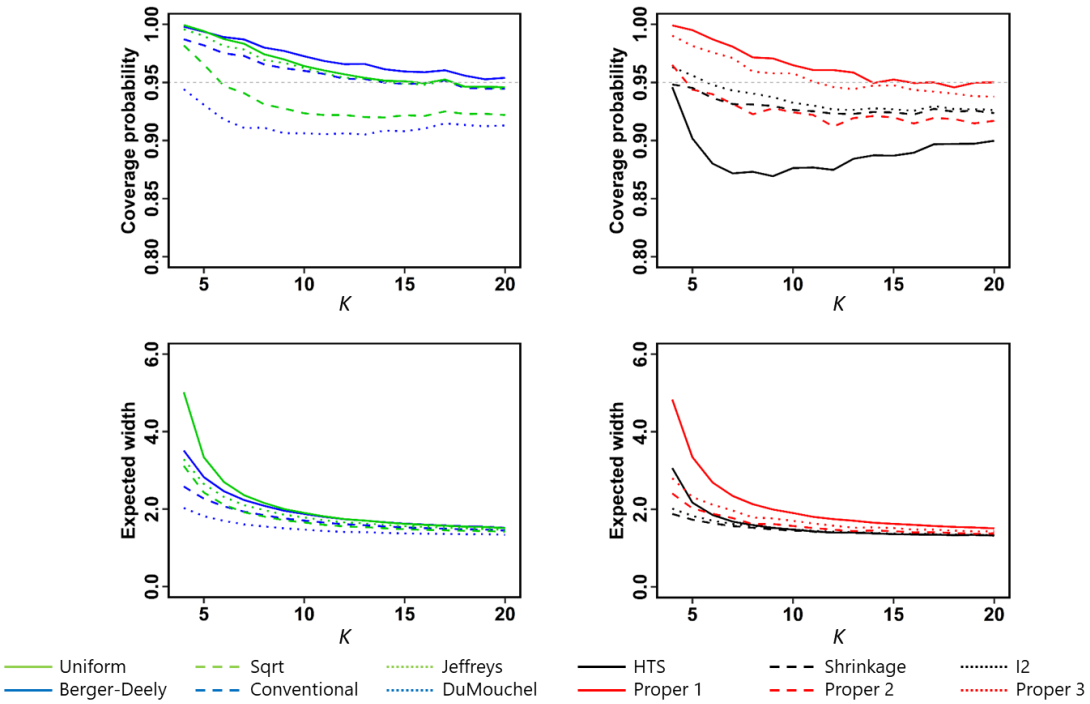

図 3.3　試験間分散を 0.10 に固定したシミュレーション実験における被覆確率および期待区間幅

3.4.5 シナリオ 4 ($\tau^2 = 0.20$)

試験間分散を 0.20 に固定したシナリオの結果を図 3.4 に示す．DuMouchel, Shrinkage，I2，Proper 2，HTS の被覆確率は一貫して 95%を下回るが，試験数が増加するに従い，被覆確率が 95%に近づいた．Sqrt，Conventional，proper 3 の被覆確率は試験数が 6 以下の場合，95%付近に位置するが，試験数が 7 以上の場合，被覆確率が減少した．ベイズ流の予測区間として，Uniform, Jeffreys，Berger−Deely，Proper 1 は試験数が 10 以下の場合，被覆確率が 95%を上回るが，試験数が 11 以上の場合，被覆確率が 95%付近に位置した．τ^2を固定したシナリオでも Proper 2 と Proper 3 の被覆確率は異なるが，試験数が増加するに従い，その差が小さくなった．期待区間幅は被覆確率の傾向を反映している．すなわち，被覆確率が高い予測区間は期待区間幅が大きくなる傾向があ

る．Uniform，Proper 1 は試験数が 4 あるいは 5 の場合，期待区間幅が著しく増加した．

τ^2を固定したシナリオ 3 および 4 の結果として，Uniform，Jeffreys，Berger–Deely，Proper 1 は試験数が 12 以上の場合，正確な頻度論的な性能を示すと考えられる．また，試験数が 10 以下の場合は，いずれのベイズ流の予測区間も好ましい頻度論的な性能を示さないと考えられる．

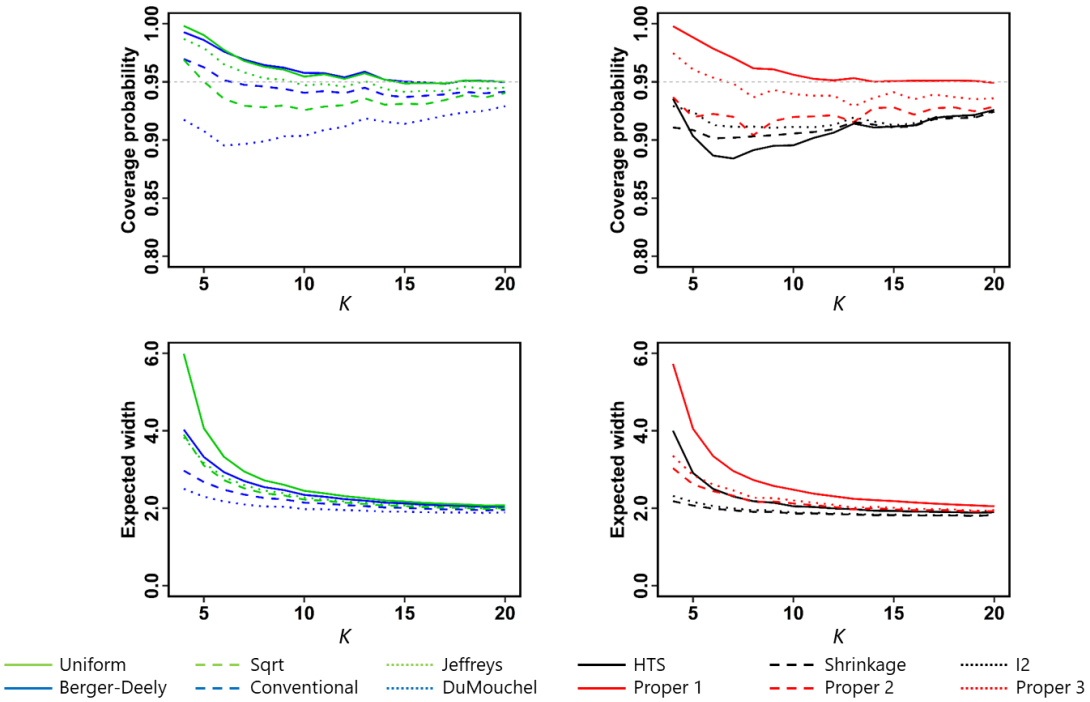

図 3.4　試験間分散を 0.20 に固定したシミュレーション実験における被覆確率および期待区間幅

3.5 事例解析

3.5.1 評価対象

ベイズ流の予測区間の特性を実際のメタアナリシスの事例で評価するために，3.3.2 項で説明した 11 種類のベイズ流の無情報事前分布を使用した予測区

間および頻度論的な手法である HTS 法に加え，Partlett and Riley (2017)により提案された HTS 法を元にした HTS-HK 法，HTS-SJ 法を使用した．また，Nagashima et al. (2019)により提案された信頼分布を使用したパラメトリックブートストラップ法に基づく pimeta 法を使用した．

HTS-HK 法

HTS 法を元にして，平均治療効果および試験間分散の制限付き最尤推定法（restricted maximum likelihood estimation; REML）による推定量および Hartung-Knapp 分散推定量を使用した予測区間である HTS-HK 法は以下のように定義される．

$$\left[\hat{\mu}_R - t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_R + \widehat{Var}_{HK}[\hat{\mu}_R]}, \hat{\mu}_R + t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_R + \widehat{Var}_{HK}[\hat{\mu}_R]}\right]$$

$$\widehat{Var}_{HK}[\hat{\mu}_R] = \frac{1}{K-1}\sum_{k=1}^{K}\frac{\hat{w}_{R,k}(Y_k - \hat{\mu}_R)^2}{\sum_{l=1}^{K}\hat{w}_{R,l}}$$

試験間分散の REML 推定量は以下の式による反復計算により算出される．初期値として DerSimonian-Laird 推定量等が使用される．

$$\hat{\tau}^2_R = \frac{\sum_{k=1}^{K}\hat{w}^2_{R,k}\left\{(Y_k - \hat{\mu}_R)^2 + 1/\sum_{l=1}^{K}\hat{w}_{R,l} - \sigma^2_k\right\}}{\sum_{k=1}^{K}\hat{w}^2_{R,k}}$$

$$\hat{w}_{R,k} = (\sigma^2_k + \hat{\tau}^2_R)^{-1}$$

$$\hat{\mu}_R = \frac{\sum_{k=1}^{K}\hat{w}_{R,k}\,Y_k}{\sum_{k=1}^{K}\hat{w}_{R,k}}$$

HTS-SJ 法

HTS 法を元にして，平均治療効果および試験間分散の REML による推定量および Sidik-Jonkman バイアス補正分散推定量を使用した予測区間である HTS-SJ 法は以下のように定義される．

$$\left[\hat{\mu}_R - t_{K-2}^{\alpha}\sqrt{\hat{\tau}_R^2 + \widehat{Var}_{SJ}[\hat{\mu}_R]}, \hat{\mu}_R + t_{K-2}^{\alpha}\sqrt{\hat{\tau}_R^2 + \widehat{Var}_{SJ}[\hat{\mu}_R]}\right]$$

$$\widehat{Var}_{SJ}[\hat{\mu}_R] = \frac{\sum_{k=1}^{K} \hat{w}_{R,k}^2 \left(1 - \hat{h}_k\right)^{-1} (Y_k - \hat{\mu}_R)^2}{\left(\sum_{k=1}^{K} \hat{w}_{R,k}\right)^2}$$

$$\hat{h}_k = \frac{2\hat{w}_{R,k}}{\sum_{l=1}^{K} \hat{w}_{R,l}} - \frac{\sum_{l=1}^{K} \hat{w}_{R,l}^2 (\sigma_k^2 + \hat{\tau}_R^2)}{(\sigma_k^2 + \hat{\tau}_R^2)\left(\sum_{l=1}^{K} \hat{w}_{R,l}^2\right)^2}$$

pimeta 法

HTS 法で使用されている大標本近似による 2 種類の仮定は，医学研究のメタアナリシスに使用される大半の事例では成立しない．そこで，それらの仮定を改善した予測区間の算出方法を Nagashima et al. (2019)が提案した．

$$\hat{\theta}_{new} = \hat{\mu} + Z\hat{\tau}_{UDL} - t_{K-1}\sqrt{\widehat{Var}_H[\hat{\mu}]}$$

$$Z = \frac{\theta_{new} - \mu}{\tau} \sim N(0,1)$$

$$\hat{\tau}_{UDL} = \frac{\sum_{k=1}^{K} \sigma_k^{-2} (Y_k - \bar{Y})^2 - (K-1)}{\sum_{k=1}^{K} \sigma_k^{-2} + \sum_{k=1}^{K} (\sigma_k^{-2})^2 / \sum_{k=1}^{K} \sigma_k^{-2}}$$

$$\bar{Y} = \frac{\sum_{k=1}^{K} \sigma_k^{-2} Y_k}{\sum_{k=1}^{K} \sigma_k^{-2}}$$

$$\widehat{Var}_H[\hat{\mu}] = \frac{1}{K-1} \sum_{k=1}^{K} \frac{\hat{w}_k}{\hat{w}_+} (Y_k - \hat{\mu})^2$$

$$\hat{w}_k = (\sigma_k^2 + \hat{\tau}^2)^{-1}$$

$$\hat{w}_+ = \sum_{k=1}^{K} \hat{w}_k$$

Z，$\hat{\tau}_{UDL}$，t_{K-1}が確率変数であり，パラメトリックブートストラップ法を利用した．また，$\hat{\tau}_{UDL}$の正確な分布については信頼分布を利用し，頻度論の枠組みでパラメータの分布推定量を構成した．

3.5.2 ジペプチジルペプチダーゼ 4（DPP-4）阻害薬の臨床試験

この事例は，2 型糖尿病患者に対するジペプチジルペプチダーゼ 4（DPP-4）阻害薬とスルホニル尿素薬を併用する低血糖のリスクを評価した事例である (Salvo et al., 2016)．対照はプラセボとスルホニル尿素薬の併用とした．評価項目は低血糖の発生であり，効果指標はリスク比である．統合する試験数は 10 であり，異質性の指標として$I^2 = 19.0\%$であった．

この事例における予測区間の結果を図 3.5 に示す．平均治療効果であるリスク比の推定値は 1.513，95%信頼区間の下限が 1.219，上限が 1.878 であった．Berger–Deely，Conventional，Jeffreys，Uniform，Proper 1，Proper 3 は他の予測区間と比較して区間が長く，HTS-HK，HTS-SJ は短かった．ベイズ流の予測区間，HTS，pimeta は予測区間に 1 を含み，HTS-HK，HTS-SJ は 1 を含まないため，予測区間の種類により異なった解釈を導く可能性が示唆される．

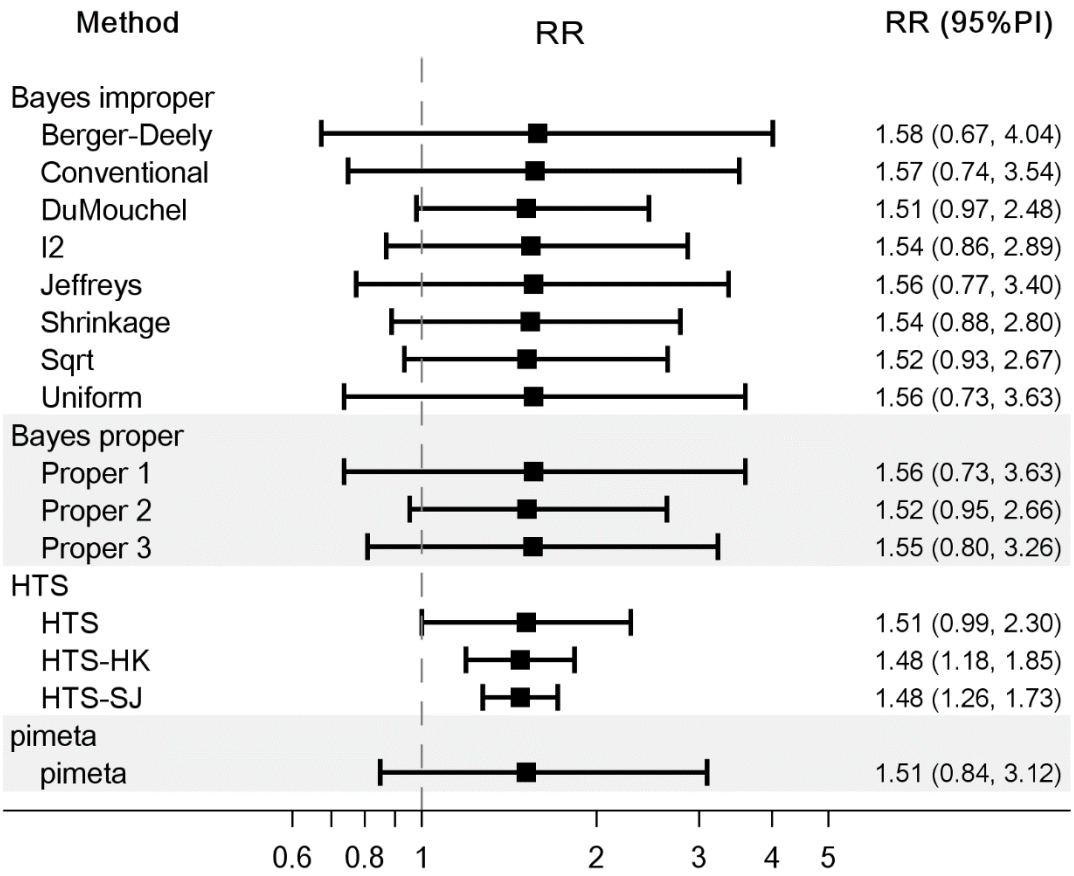

図 3.5　DPP-4 阻害薬の臨床試験における予測区間の結果

3.5.3 抗うつ薬の臨床試験

この事例は，線維筋痛症患者の苦痛に対する抗うつ薬の治療効果を評価した事例である(Häuser et al., 2009)．評価項目は Visual analog scale (VAS) 等に基づく痛みの質問であり，効果指標は標準化平均差である．統合する試験数は 22 であり，異質性の指標として$I^2 = 44.9\%$であった．

この事例における予測区間の結果を図 3.6 に示す．ベイズ流の予測区間では，Berger−Deely，Conventional，Jeffreys，Uniform，Proper 1，Proper 3 は予測区間に 0 を含み，DuMouchel，I2，Shrinkage，Sqrt，Proper 2 は予測区間に 0 を含まなかった．頻度論的手法の予測区間では，HTS，HTS-HK，HTS-SJ は予測

区間に 0 を含まないが，pimeta は 0 を含んだ．それゆえ，選択した事前分布の種類により結果に対する解釈が異なる可能性が示唆される．

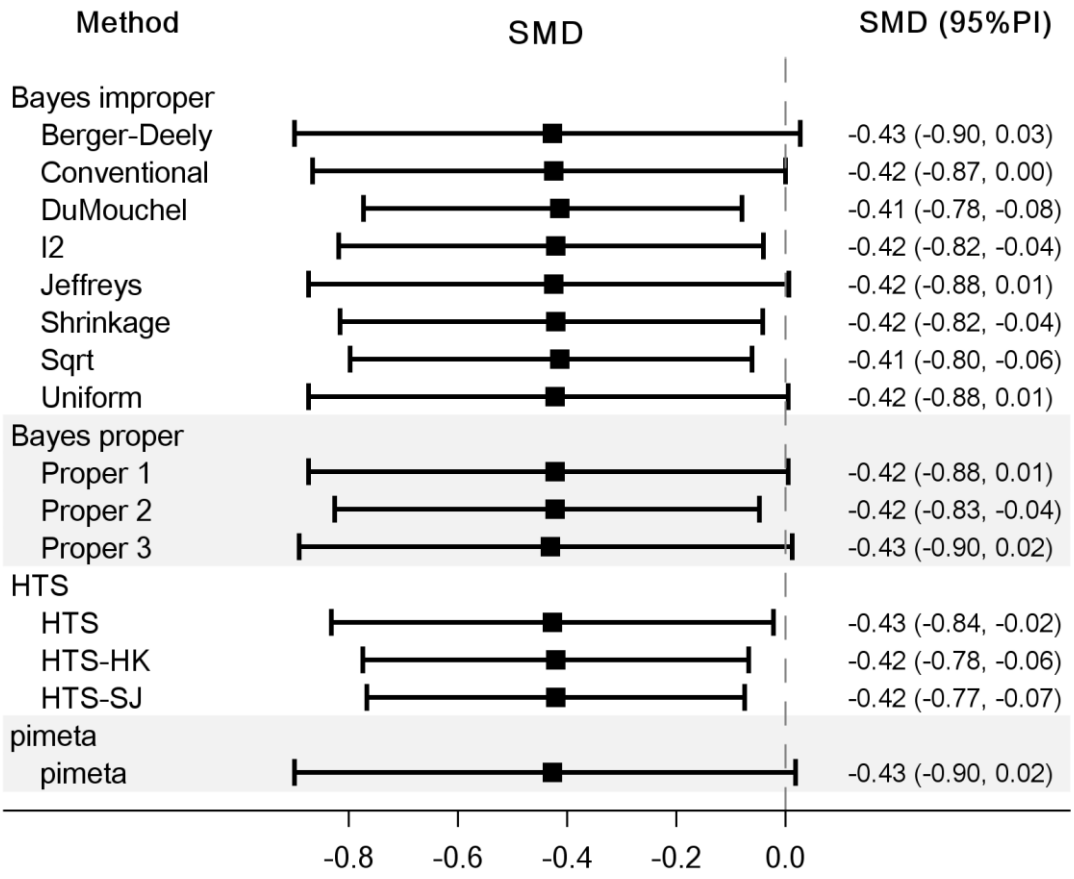

図 3.6　抗うつ薬の臨床試験における予測区間の結果

3.6 考察

予測区間は対象集団における治療効果の異質性と不確実性を評価する有用な手法である．ベイズ流の手法は一般的な予測問題に対して実践で有用なアプローチであるが，本章におけるシミュレーション実験の結果によるとその予測区間の頻度論的な性能は必ずしも正確ではなく，実践で慎重に扱われるべきであることが示唆された．ベイズ流予測区間の被覆確率は特に試験間分散が小さい

場合，名目水準の 95%を上回っており，これはベイズ推定により試験間分散の事後分布が高値に分布していることで予測区間が長くなったことが原因であると考えられる．また，試験数が増加するに従い，ベイズ流予測区間の被覆確率は 95%に近づく傾向がある．これは，事前分布の相対的な情報量が観測値と比較して低下するためと考えられ，試験数が 10 以下の場合，被覆確率が 95%であるベイズ流予測区間は存在しなかった．各予測区間については，Uniform, Berger-Deely，Proper 1 はいずれの条件においても被覆確率が 95%以上であり，試験間分散が 0.2 で試験数が 11 以上の場合は被覆確率が 95%になる．これらの予測区間を使用することで一貫して保守的な結果が得られるが，事前分布がτに対する一様分布である Uniform， Proper 1 については，試験数が 4 および 5 と少ない場合，期待区間幅が大幅に上昇することで過剰に保守的になることに注意が必要である．その他のベイズ流予測区間は，被覆確率が 95%を上回る場合と下回る場合が存在し， Sqrt，DuMouchel，Shrinkage，I2，Proper 2 は，条件により被覆確率が 90%まで低下する場合がある．また，Proper 2 および Proper 3 はいずれも事前分布が逆ガンマ分布であるが被覆性能は異なり，Proper 3 が Proper 2 より被覆確率が高くなっている．これは，Proper 3 の事前分布の方が高値に分布していることに起因していると考えられる．このようにあらゆる条件下において被覆性能が一貫して妥当な予測区間は本章で評価した 11 種類の事前分布において存在しなかった．一方，比較対照とした HTS 法による頻度論的な予測区間は，いずれの条件においても被覆確率は 95%以下で過小評価されており，条件により 85%を下回ることもある．この結果は Partlett and Riley (2017)や Nagashima et al. (2019)と同様の結果であり，大標本近似が成立しないことが主要な原因であると考えられる． また， HTS 法の期待区間幅は，試験間分散が小さい場合，ベイズ流予測区間と比較して短くなり，DerSimonian-Laird 推定量が過小評価されることが原因であると考えられる．

頻度論的な予測区間の HTS 法に使用される試験間分散に関しては，これまでに様々な推定量が提案されている．Partlett and Riley (2017)はシミュレーション実験で試験間分散の推定量として従来の DerSimonian-Laird 推定量や REML 推定量等の性能を評価した結果，REML 推定量の性能が優れていることを示している．本章では，実践における頻度論的な予測区間の性能をベイズ流の予測区間と比較するため，実践で最も使用されている DerSimonian-Laird 推定量を HTS 法の試験間分散の推定量として適用した．

以上の結果により，これまで実質的に頻度論的手法の代替手段として使用されてきたベイズ流の予測区間は，現実的な条件下で，多くの場合，頻度論的予測区間としては解釈することができないと考えられる．実践で用いる上では，Uniform，Berger-Deely，Proper 1 を使用することで一貫して保守的な結果が得られることが期待される．また，ベイズ流予測区間の事前分布を任意に選択することで恣意的な結果が得られる可能性があるため，メタアナリシスを実施する前に事前分布を特定することが望まれる． 一方，頻度論的な予測区間として， Nagashima et al. (2019)が提案された pimeta 法は本章のシミュレーション実験で評価しなかったものの HTS 法と比較して被覆性能が向上した予測区間を構築できるため，HTS 法の代替手段の一つになると考えられる．

今後，単変量メタアナリシス以外のメタアナリシスに対するベイズ流予測区間の評価が必要であると考えられる．診断法の性能を評価する 2 変量メタアナリシスにおいては，診断法の感度および特異度を評価する予測区間に相当する予測楕円（prediction ellipse）の使用が推奨されており(Reitsma et al., 2005)，多変量メタアナリシスおいては，多変量正規分布を仮定したモデルより予測区間を算出することができる(Snell et al., 2016)．ネットワークメタアナリシスにおいても，試験間の異質性の問題があることで予測区間の重要性が認識されており，統計ソフトウェアのパッケージで予測区間を算出できるものの，予測区間

を掲載している事例は主要な医学雑誌において 9%と少ないことが報告されている(Lin, 2019). いずれの種類のメタアナリシスにおいても試験間の異質性の評価は重要であり，予測区間の使用が普及するものと想定されるが，事前分布の種類により予測区間の頻度論的な性能を定量的に評価したエビデンスはこれまで存在しない．本章で実施したシミュレーション実験の方法を拡張することで評価できると想定されるため，今後の研究課題の一つとして挙げられる．

第 4 章 ベイズ流手法による外れ値試験の検出と影響力解析

4.1 背景と目的

メタアナリシスでは治療効果の試験間の異質性が大きい事例が存在することがあり，そのような事例では他の試験と比較して極端な値を示す外れ値（outlier）とされる試験が含まれる可能性がある．外れ値が存在する場合，変量効果モデルで適切にモデル化できないことが試験間の異質性や平均治療効果の推定に大きく影響を及ぼす可能性がある．外れ値への対処法として，外れ値の影響を低減するアプローチと，外れ値の検出および影響力を解析するアプローチが考えられ，前者では変量効果の分布に裾の重い分布や歪んだ分布を仮定する手法が提案されているが(Baker and Jackson, 2008; Baker and Jackson, 2016; Beath, 2014; Lee and Thompson, 2008)，システマティックレビューにおいては外れ値を解析対象から除外した感度分析が推奨されており(Higgins et al., 2019)，後者のアプローチはメタアナリシスの実践において重要である．外れ値の検出法として，古くは，要約統計量をプロットしたグラフによる視覚的な評価が使用されており(Cooper et al., 2009; Julious and Whitehead, 2012; Wang and Bushman, 1998)，簡便に評価できる利点はあるものの評価が主観的であることが課題であった．近年では，回帰分析における影響力を診断する頻度論的な手法が変量効果モデルのメタアナリシスに適用されている(Gumedze and Jackson, 2011; Viechtbauer and Cheung, 2010)．また，ベイズ流のメタアナリシスは，試験間分散の不確実性を評価できる利点があり，頻度論的なメタアナリシスの代替法として近年急速に普及している．しかしながら，ベイズ流の単変量メタアナリシスにおいては，外れ値の検出および影響力を診断する簡便な手法が提案されておらず，実務家がメタアナリシスの実践で利用することはできなかった．

これまで，Zhang et al. (2015) は，ネットワークメタアナリシスにおけるベイズ流の外れ値の検出および影響力を診断する方法として，相対距離，標準化残差，ベイズ流 P 値，混合尺度正規分布における尺度パラメータの事後推定値を提案しており，Matsushima et al. (2020) は，診断法のメタアナリシスにおいて，感度と偽陽性率の 2 変量を統合した指標として，ベイズ流の相対距離，標準化残差，ベイズ流 P 値，および要約 ROC (Receiver operating characteristics)曲線下面積を提案している．本章では，ベイズ流の単変量メタアナリシスにおける相対距離，標準化残差，ベイズ流 P 値，および尺度混合正規分布における尺度パラメータの事後推定値の指標を開発し，外れ値の検出および影響力診断の特性を，慢性腰痛患者に対する治療法(Rubinstein et al., 2019)，安定冠動脈疾患患者に対する治療薬(Bangalore et al., 2017)，妊娠糖尿病の既往の 2 型糖尿病発症リスク(Vounzoulaki et al., 2020)，および新生児の呼吸窮迫症候群発症に対する出産前ステロイド投与(Saccone and Berghella, 2016)の 4 つの事例解析を通じて確認した．

4.2 ベイズ流の手法

4.2.1 相対距離

回帰分析では外れ値を検出するために leave-one-out 交差検証の手法を用いた指標が提案されており，その一つとして Cook の距離に類似した相対距離が挙げられる．単変量メタアナリシスにおける変量効果モデルでの相対距離を以下のように定義する．

$$RD_k = \left| \frac{E_{\mu|\boldsymbol{y}}(\mu|\boldsymbol{y}) - E_{\mu|\boldsymbol{y}_{(k)}}(\mu|\boldsymbol{y}_{(k)})}{E_{\mu|\boldsymbol{y}}(\mu|\boldsymbol{y})} \right|$$

$\boldsymbol{y_{(k)}}$は全ての試験から試験kを除外したデータを表す．RD_kは，平均治療効果の事後分布における平均が試験kを除外した場合にどの程度変化するかを表す．RD_kは，メタアナリシスにおいて最も関心のある平均治療効果に対する各試験の影響が直接反映される利点がある．一方，RD_kは分母が 0 付近になることが期待される場合に不安定であり，メタアナリシスにおいては効果指標が標準化平均差である場合，結果の解釈に注意が必要である．外れ値の閾値に関して，回帰分析における Cook の距離では自由度を観測数および観測数と説明変数の差としたF分布に従うと仮定し，F分布の中央値を閾値とすることが提案されているが(Cook, 1977)，本章ではRD_kの閾値を仮定せず，高値である試験を平均治療効果への影響が大きい試験として扱うこととした．

4.2.2 標準化残差

相対距離と同様に leave-one-out 交差検証の手法を用いた指標である標準化残差は，ベイズ流の枠組みにおいても広く用いられている(Carlin and Louis, 2009)．これまで，単変量メタアナリシスにおける頻度論的な手法としてスチューデント化残差が Viechtbauer and Cheung (2010) により提案されていたが，ベイズ流のメタアナリシスに適用できる指標は提案されていなかった．単変量メタアナリシスにおける変量効果モデルでの標準化残差を以下のように定義する．

$$SR_k = \frac{y_k - E_{\theta_{new}|\boldsymbol{y_{(k)}}}(\theta_{new}|\boldsymbol{y_{(k)}})}{\sqrt{Var_{\theta_{new}|\boldsymbol{y_{(k)}}}(\theta_{new}|\boldsymbol{y_{(k)}})}}$$

y_kは試験kの治療効果の推定値，θ_{new}は仮定したモデルにおいて新たな試験を実施した際に想定される治療効果を表す．SR_kは，試験kを除外したデータを使用して算出した新たな試験の治療効果の事後予測分布における平均および分散を使用して算出する．SR_kはその絶対値が高値であるほど試験kが外れ値である可能性が高いと想定される．その閾値は，標準正規分布を参照して上側およ

び下側 2.5%分位点とし，SR_kが 1.96 より高値または-1.96 より低値である場合，試験kを外れ値と判定することとした．

4.2.3 ベイズ流 P 値

Leave-one-out 交差検証とは異なる手法として，従来から事後予測モデルチェックに基づく手法が提案されている(Rubin, 1984)．事後予測モデルチェックは観測されたデータと階層ベイズモデルからサンプリングされた事後予測データとの不一致の程度を評価するものである．ベイズ流 P 値はこの手法に基づく指標であり，単変量メタアナリシスにおける変量効果モデルでの試験kの不一致の程度を表す指標を以下のように定義する．

$$D_k(y_k, \boldsymbol{\xi}) = \frac{\left[y_k - E_{\theta_{new}|\xi}(\theta_{new}|\boldsymbol{\xi})\right]^2}{Var_{\theta_{new}|\xi}(\theta_{new}|\boldsymbol{\xi})}$$

$$D_{k,new}(y_{k,new}, \boldsymbol{\xi}) = \frac{\left[y_{k,new} - E_{\theta_{new}|\xi}(\theta_{new}|\boldsymbol{\xi})\right]^2}{Var_{\theta_{new}|\xi}(\theta_{new}|\boldsymbol{\xi})}$$

$y_{k,new}$は仮説的にサンプリングした新たな試験kの治療効果の推定値を表す．$D_k(y_k, \boldsymbol{\xi})$はカイ二乗統計量に類似しており，試験$k$の治療効果の推定値と，新たな試験の治療効果の事後予測分布の平均の差の二乗をその分散で割った指標である．$D_{k,new}(y_{k,new}, \boldsymbol{\xi})$は，$D_k(y_k, \boldsymbol{\xi})$における試験$k$の治療効果の推定値を事後予測分布からサンプリングされた新たな試験kの治療効果の推定値に置き換えた指標である．本章では 2.2 節に記載の通り，事後分布を算出するためにマルコフ連鎖モンテカルロ法（MCMC）を使用しており，$y_{k,new}$は平均を MCMC でサンプリングされたθ_{new}，分散を試験kの試験内分散とした正規分布からの事後予測サンプルである．$D_k(y_k, \boldsymbol{\xi})$と$D_{k,new}(y_{k,new}, \boldsymbol{\xi})$からベイズ流 P 値を以下のように定義する．

$$P_{D_k} = P\left[D_k(y_k, \boldsymbol{\xi}) < D_{k,new}\left(y_{k,new}, \boldsymbol{\xi}\right)|\boldsymbol{y}\right]$$

$$= \int P\left\{D_k(y_k, \boldsymbol{\xi}) < D_{k,new}\left(y_{k,new}, \boldsymbol{\xi}\right)|\boldsymbol{\xi}\right\} p(\boldsymbol{\xi}|\boldsymbol{y}) d\boldsymbol{\xi}$$

P_{D_k}は，MCMC を使用する場合，サンプリングした$D_{k,new}\left(y_{k,new}, \boldsymbol{\xi}\right)$が$D_k(y_k, \boldsymbol{\xi})$を上回った割合として算出される．$P_{D_k}$の閾値を一般的な有意水準である 5%とし，閾値を下回る場合，試験kを外れ値と判定することとした．

4.2.4 尺度混合正規分布における尺度パラメータの事後推定値

尺度混合正規分布（scale mixtures of normals）は変量効果が様々な分布に従うと仮定するモデルである(Carlin and Louis, 2009)．単変量メタアナリシスにおける尺度混合正規分布を以下のように定義する．

model 1　$y_k \sim N(\theta_k, \lambda_k \sigma_k^2)$,　$\theta_k \sim N(\mu, \tau^2)$,　$\lambda_k = 1$

model 2　$y_k \sim N(\theta_k, \lambda_k \sigma_k^2)$,　$\theta_k \sim N(\mu, \tau^2)$,　$\lambda_k \sim Exp(2)$

model 3　$y_k \sim N(\theta_k, \lambda_k \sigma_k^2)$,　$\theta_k \sim N(\mu, \tau^2)$,　$\lambda_k \sim InvGamma(1,1)$

変量効果モデルは試験内分散が正規分布に従うと仮定する．一方，尺度混合正規分布では試験内分散を試験kの尺度パラメータであるλ_kにより調整し，λ_kが 1 である場合（model 1），変量効果モデルと同様にy_kは正規分布に従うが，λ_kがパラメータ 2 の指数分布に従う場合（model 2），y_kは二重指数分布に従い，λ_kがパラメータ(1, 1)の逆ガンマ分布に従う場合（model 3），y_kは自由度 2 の t 分布に従う．二重指数分布や t 分布のような正規分布より裾の重い分布を仮定し，それぞれの分布に対するy_kの当てはまりを評価することで外れ値を検出することができる．すなわち，裾の重い分布を仮定した尺度混合正規分布のλ_kの事後分布が高値である場合，試験kは外れ値である可能性が高いと想定される．二重指数分布を仮定した model 2 および自由度 2 の t 分布を仮定した model 3 におけるλ_kの閾値として，正規分布に従うことを否定することが妥当

な状況としてλ_kが 1 以上となる確率が 95%を上回る場合に試験kを外れ値と判定することとした.

4.3 事例解析

本節の事例解析では，ベイズ流の枠組みのもとでパラメータの事後分布を推定するためにマルコフ連鎖モンテカルロ法（MCMC）を用いた(Lunn et al., 2009)．事前分布を，$\mu \sim N(0,10000)$, $\tau^2 \sim InvGamma(0.01,0.01)$と仮定した.

4.3.1 慢性腰痛患者に対する治療法

この事例は，慢性腰痛患者に対する脊椎手技療法の有効性を評価したメタアナリシスである(Rubinstein et al., 2019)．脊椎手技療法とは，慢性腰痛を持つ患者に対して医療従事者が実施する治療法であり，患者の脊椎を可動域内で動かしたり，可動域の限界において関節を強く押す治療法である．これまで，脊椎手技療法の有効性に対する見解が分かれており，原著論文では脊椎手技療法の有効性を運動療法や投薬といった慢性腰痛に対する一般的な治療と比較した．対象としたメタアナリシスにおける評価項目は，治療して一ヶ月後の痛みの強度に関する VAS（Visual Analogue Scale）や ODI（Oswestry Disability Index）等を 100 点換算したスコアである．脊椎手技療法と一般的な治療を比較した 23 の試験に対するベイズ流メタアナリシスの結果として，効果指標を平均差とした場合の結果を図 4.1 に示す．効果指標が平均差である場合，試験間分散の事後平均は114.65であり，異質性の高い試験群であった．実際に，試験 16，試験 18，試験 23 の治療効果の推定値は平均治療効果の推定値から視覚的に離れていた．平均治療効果の事後平均は −3.18（95 %信用区間：−7.97 ~ 1.61）であった.

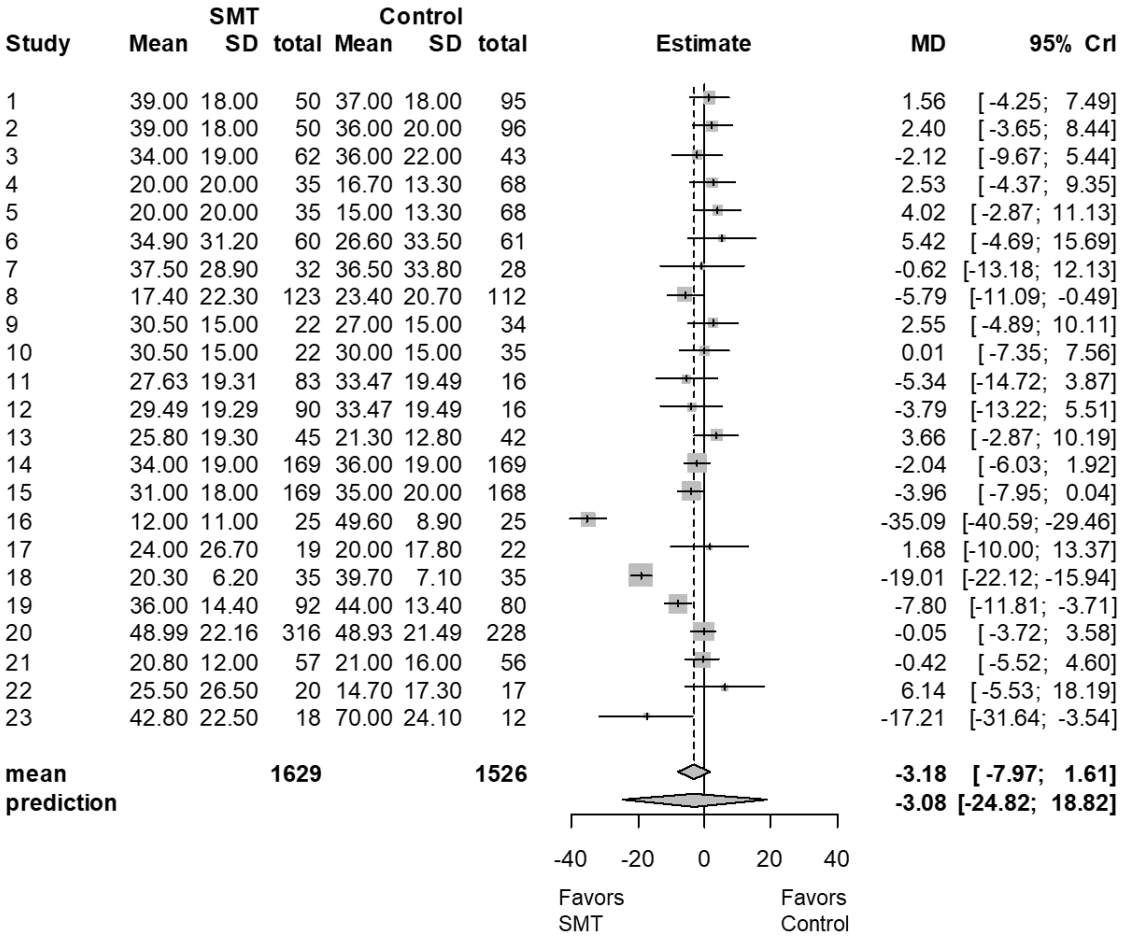

図 4.1　慢性腰痛患者に対する治療法のベイズ流メタアナリシスにおける試験毎の平均差についてのフォレストプロット

試験毎の相対距離，標準化残差，ベイズ流 P 値を表 4.1 に，尺度混合正規分布における尺度パラメータの事後推定値を表 4.2 に示す．相対距離は，試験 16 が 0.490 と最も高く，試験 18 が 0.265，試験 23 が 0.222 と比較的高かった．標準化残差により外れ値として検出された試験は，試験 16 および試験 23 であった．また，ベイズ流 P 値により外れ値として検出された試験は，試験 16 のみであった．尺度混合正規分布における尺度パラメータの事後推定値により外れ値として検出された試験は，二重指数分布を仮定した場合は存在せず，自由度 2 の *t* 分布を仮定した場合は試験 16，試験 18，試験 23 であった．閾値のない相

対距離以外の指標で共通して外れ値と検出されたのは試験 16 であり，試験 18 および試験 23 が検出されるかどうかは指標により異なった.

表 4.1　効果指標を平均差とした慢性腰痛患者に対する治療法のベイズ流メタアナリシスにおける相対距離, 標準化残差, およびベイズ流 P 値

試験	相対距離	標準化残差	ベイズ流 P 値
1	0.078	0.485	0.639
2	0.093	0.580	0.580
3	0.015	0.111	0.923
4	0.095	0.608	0.570
5	0.121	0.772	0.469
6	0.145	1.088	0.348
7	0.042	0.388	0.764
8	0.047	-0.261	0.787
9	0.096	0.627	0.561
10	0.052	0.343	0.751
11	0.037	-0.245	0.811
12	0.013	-0.073	0.940
13	0.115	0.725	0.500
14	0.017	0.111	0.921
15	0.016	-0.075	0.931
16	0.490	-5.254	0.005
17	0.080	0.669	0.583
18	0.265	-1.634	0.137
19	0.080	-0.449	0.645
20	0.051	0.303	0.768
21	0.045	0.279	0.787
22	0.156	1.329	0.285
23	0.222	-2.340	0.086

表 4.2　効果指標を平均差とした慢性腰痛患者に対する治療法のベイズ流メタアナリシスの尺度混合正規分布における尺度パラメータの事後推定値

試験	二重指数分布		*t* 分布	
	事後平均 [95%信用区間]	$\Pr(\lambda_k \geqq 1)$	事後平均 [95%信用区間]	$\Pr(\lambda_k \geqq 1)$
1	1.88 [0.04, 7.15]	0.584	2.79 [0.26, 13.15]	0.550
2	1.89 [0.05, 7.06]	0.587	2.97 [0.29, 15.37]	0.610
3	1.81 [0.04, 6.85]	0.567	2.28 [0.24, 11.21]	0.481
4	1.86 [0.04, 7.04]	0.579	3.05 [0.28, 13.95]	0.589
5	1.95 [0.05, 7.12]	0.603	4.19 [0.34, 18.70]	0.690
6	1.98 [0.06, 6.87]	0.621	4.04 [0.38, 19.11]	0.722
7	1.59 [0.04, 6.20]	0.514	2.12 [0.23, 10.36]	0.451
8	1.87 [0.05, 6.81]	0.587	4.77 [0.36, 23.16]	0.768
9	1.84 [0.05, 6.72]	0.584	2.74 [0.28, 13.33]	0.578
10	1.83 [0.05, 6.64]	0.582	2.41 [0.24, 11.25]	0.486
11	1.72 [0.04, 6.44]	0.551	3.00 [0.29, 13.96]	0.590
12	1.68 [0.04, 6.29]	0.548	2.60 [0.25, 11.51]	0.521
13	1.90 [0.05, 7.04]	0.590	3.74 [0.32, 19.08]	0.688
14	1.89 [0.05, 6.95]	0.585	2.85 [0.26, 12.61]	0.547
15	1.93 [0.05, 7.17]	0.596	4.43 [0.32, 21.33]	0.702
16	5.37 [0.19, 15.19]	0.871	155.78 [16.45, 771.98]	1.000
17	1.69 [0.04, 6.26]	0.551	2.56 [0.25, 10.75]	0.508
18	2.39 [0.06, 8.85]	0.655	123.45 [3.16, 572.41]	0.990
19	1.93 [0.04, 7.12]	0.594	11.28 [0.46, 58.67]	0.904
20	1.95 [0.06, 7.07]	0.606	2.64 [0.25, 13.19]	0.530
21	1.89 [0.05, 7.09]	0.588	2.34 [0.24, 11.33]	0.491
22	2.05 [0.06, 7.20]	0.638	4.19 [0.41, 19.68]	0.751
23	2.93 [0.19, 8.63]	0.824	10.73 [1.08, 51.41]	0.980

また，脊椎手技療法と一般的な治療を比較した 23 の試験に対するベイズ流メタアナリシスの結果として，効果指標を標準化平均差とした場合の結果を図 4.2 に示す．試験間分散の事後平均は0.80であり，試験 16，試験 18 の治療効果の推定値は平均治療効果の推定値から視覚的に離れていた．平均治療効果の事後平均は −0.27（95 %信用区間：−0.67 ~ 0.10）であった．

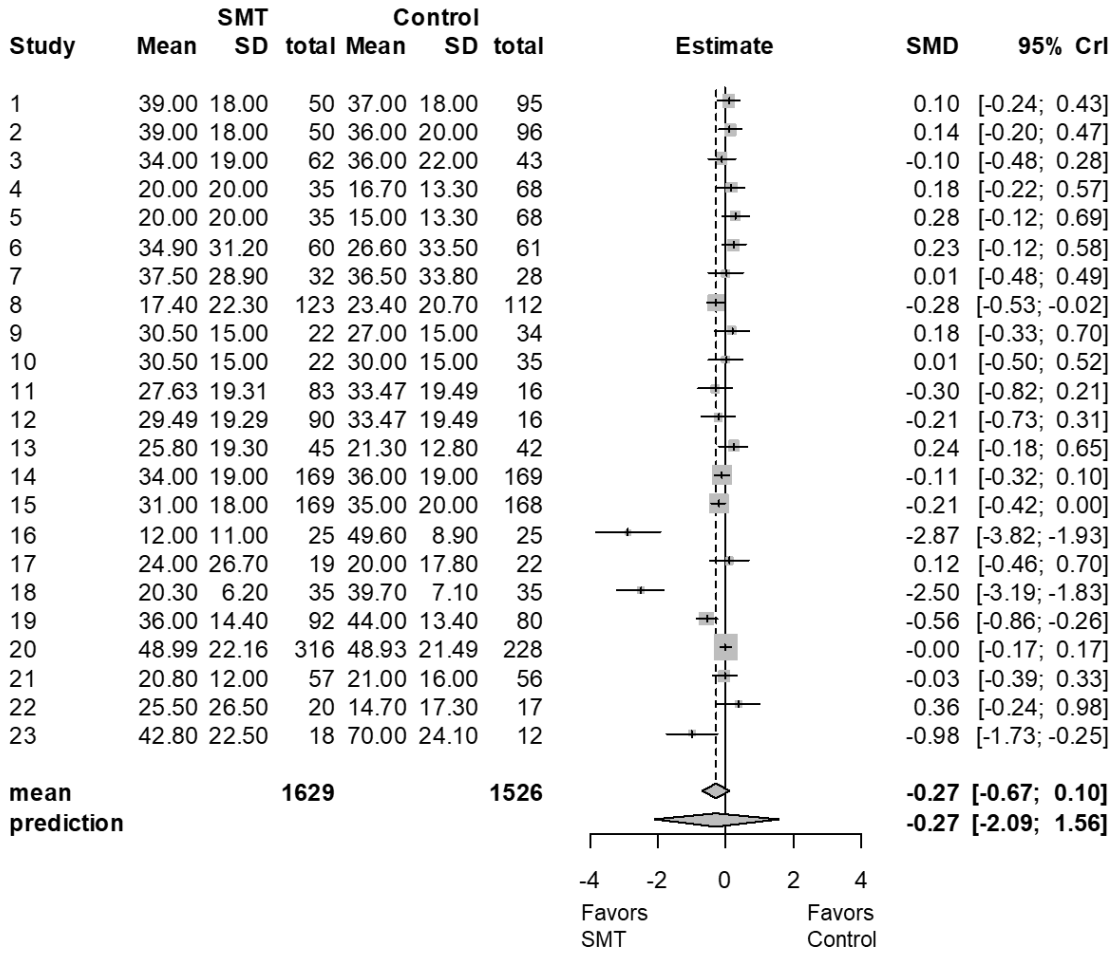

図 4.2　慢性腰痛患者に対する治療法のベイズ流メタアナリシスにおける試験毎の標準化平均差についてのフォレストプロット

試験毎の相対距離，標準化残差，ベイズ流 P 値を表 4.3 に，尺度混合正規分布における尺度パラメータの事後推定値を表 4.4 に示す．相対距離は，試験 16 が 0.510，試験 18 が 0.484 と高かった．標準化残差およびベイズ流 P 値により外れ値として検出された試験は，いずれも試験 16 および試験 18 であった．尺度混合正規分布における尺度パラメータの事後推定値により外れ値として検出された試験は，二重指数分布と自由度 2 の *t* 分布を仮定したいずれの場合も試験 16，試験 18，試験 23 であった．なお，試験 16 および試験 18 はこの事例の

原著論文で，外れ値として検出された試験の内，バイアスリスクが高いと言及されていた.

表 4.3　効果指標を標準化平均差とした慢性腰痛患者に対する治療法のベイズ流メタアナリシスにおける相対距離, 標準化残差, およびベイズ流 P 値

試験	相対距離	標準化残差	ベイズ流 P 値
1	0.074	0.430	0.668
2	0.081	0.479	0.635
3	0.039	0.198	0.848
4	0.088	0.537	0.599
5	0.105	0.658	0.518
6	0.097	0.592	0.557
7	0.058	0.341	0.749
8	0.008	0.001	0.990
9	0.088	0.563	0.588
10	0.058	0.343	0.742
11	0.004	-0.024	0.969
12	0.020	0.081	0.945
13	0.098	0.610	0.555
14	0.039	0.191	0.851
15	0.020	0.075	0.948
16	0.510	-5.957	0.002
17	0.079	0.502	0.634
18	0.484	-4.293	0.012
19	0.043	-0.325	0.730
20	0.058	0.311	0.757
21	0.053	0.292	0.776
22	0.119	0.831	0.437
23	0.125	-0.980	0.364

表 4.4　効果指標を標準化平均差とした慢性腰痛患者に対する治療法のベイズ流メタアナリシスの尺度混合正規分布における尺度パラメータの事後推定値

試験	二重指数分布		t 分布	
	事後平均 [95%信用区間]	Pr($\lambda_k \geqq 1$)	事後平均 [95%信用区間]	Pr($\lambda_k \geqq 1$)
1	1.65 [0.04, 6.27]	0.540	2.49 [0.25, 12.50]	0.533
2	1.75 [0.05, 6.57]	0.564	2.77 [0.26, 13.67]	0.563
3	1.57 [0.03, 6.28]	0.497	2.43 [0.25, 11.69]	0.518
4	1.77 [0.05, 6.62]	0.572	3.01 [0.27, 14.43]	0.577
5	2.13 [0.09, 7.30]	0.673	3.49 [0.30, 17.83]	0.656
6	2.00 [0.07, 7.05]	0.631	3.56 [0.29, 17.95]	0.638
7	1.44 [0.03, 5.95]	0.466	2.41 [0.24, 11.47]	0.484
8	2.11 [0.07, 7.23]	0.652	4.35 [0.30, 22.07]	0.678
9	1.73 [0.06, 6.42]	0.568	2.54 [0.26, 12.53]	0.551
10	1.42 [0.03, 5.83]	0.457	2.19 [0.23, 10.89]	0.486
11	1.76 [0.06, 6.55]	0.574	2.92 [0.28, 14.16]	0.587
12	1.56 [0.04, 6.01]	0.509	2.38 [0.25, 12.14]	0.524
13	1.94 [0.08, 6.84]	0.621	3.54 [0.29, 15.96]	0.627
14	1.75 [0.04, 6.45]	0.558	3.02 [0.26, 15.05]	0.564
15	1.92 [0.05, 6.78]	0.607	3.59 [0.28, 19.30]	0.632
16	8.48 [3.95, 15.85]	1.000	61.86 [6.38, 270.72]	1.000
17	1.54 [0.04, 6.12]	0.495	2.52 [0.25, 11.47]	0.515
18	8.95 [4.17, 16.48]	1.000	66.82 [7.27, 332.11]	1.000
19	3.33 [0.24, 9.35]	0.861	10.84 [0.48, 52.95]	0.896
20	1.80 [0.05, 6.84]	0.565	3.21 [0.26, 15.63]	0.566
21	1.52 [0.03, 6.05]	0.496	2.51 [0.24, 11.47]	0.508
22	2.23 [0.17, 7.19]	0.716	4.87 [0.34, 17.78]	0.678
23	3.52 [0.81, 9.12]	0.952	9.18 [0.82, 42.77]	0.953

4.3.2 安定冠動脈疾患患者に対する治療薬

この事例は，心不全のない安定冠動脈疾患患者に対するレニン・アンジオテンシン系阻害薬の有効性を評価したメタアナリシスである(Bangalore et al., 2017)．レニン・アンジオテンシン系阻害薬は，心不全のない安定冠動脈疾患患者の心血管イベントや死亡を抑制することで，その使用をガイドラインで強く推奨されていたが，現行治療を上乗せした近年の研究ではプラセボに対する有

効性が示されていなかった．原著論文では評価項目を死亡，心血管死，心筋梗塞，狭心症，脳卒中，心不全，血行再建術，糖尿病発症，および有害事象による服薬中止とし，対照をカルシウム拮抗薬等の実薬，あるいはプラセボとしてレニン・アンジオテンシン系阻害薬の有効性を評価した．対象としたメタアナリシスは，効果指標を 1,000 人年当たりの死亡率の比とし，プラセボを対照としたレニン・アンジオテンシン系阻害薬の有効性を評価した 18 のランダム化比較試験によるものであり，この事例に対するベイズ流メタアナリシスの結果を図 4.3 に示す．試験間分散の事後平均は0.13であり，実際に試験 4 の治療効果の推定値は平均治療効果の推定値から視覚的に離れていた．平均治療効果の事後平均は 0.82（95 %信用区間：0.65 ~ 1.03）であった．

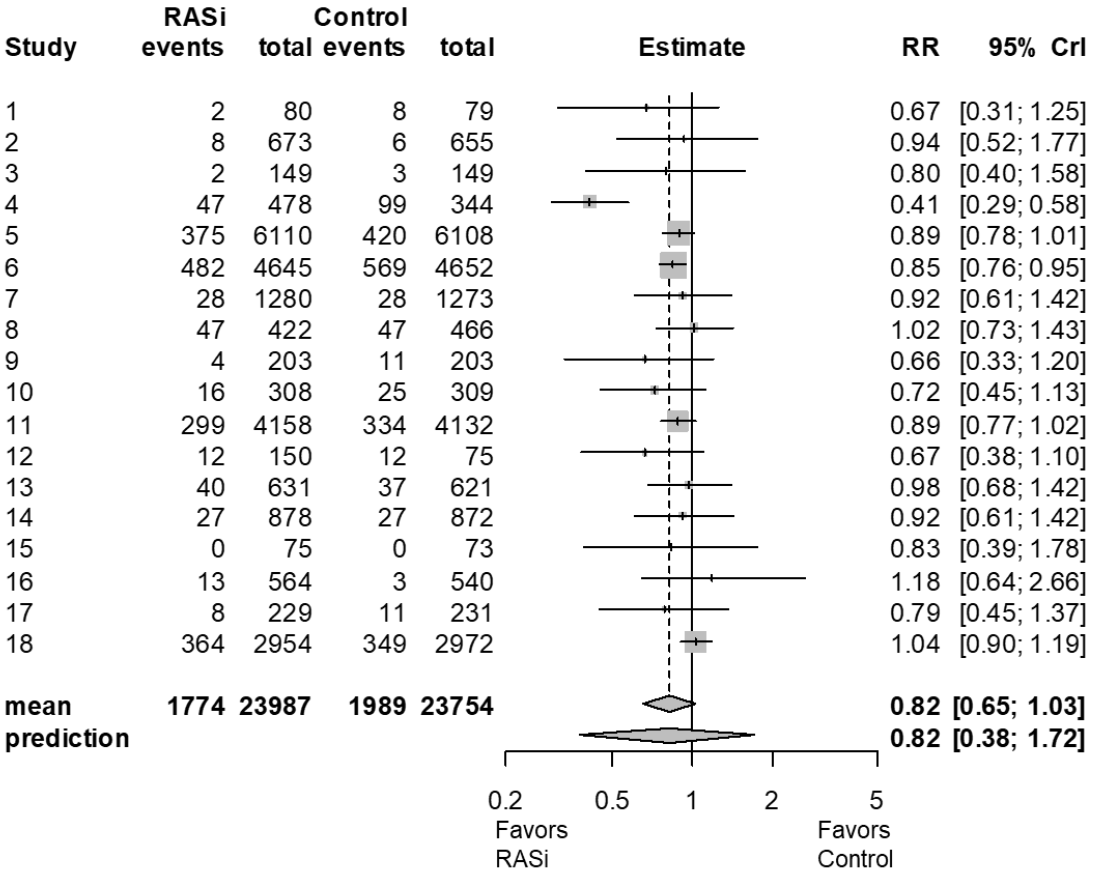

図 4.3　安定冠動脈疾患患者に対する治療薬のベイズ流メタアナリシスにおける試験毎の死亡率の比についてのフォレストプロット

この事例の試験毎の相対距離，標準化残差，ベイズ流 P 値を表 4.5 に，尺度混合正規分布における尺度パラメータの事後推定値を表 4.6 に示す．相対距離は，試験 4 が 0.534 と最も高かった．標準化残差により外れ値として検出された試験は，試験 1，試験 4，試験 9，試験 16 であった．また，ベイズ流 P 値により外れ値として検出された試験は，試験 4 および試験 16 であった．尺度混合正規分布における尺度パラメータの事後推定値により外れ値として検出された試験は，二重指数分布を仮定した場合は存在せず，自由度 2 の *t* 分布を仮定した場合は試験 4 であった．閾値のない相対距離以外の指標で共通して外れ値と検出されたのは試験 4 であり，試験 1，試験 9，試験 16 が検出されるかどうか

は指標により異なった．この事例の原著論文では，外れ値として検出された試験の内，試験 9 のバイアスリスクが不明と言及されていた．

表 4.5　安定冠動脈疾患患者に対する治療薬のベイズ流メタアナリシスにおける相対距離, 標準化残差, およびベイズ流 P 値

試験	相対距離	標準化残差	ベイズ流 P 値
1	0.124	-3.244	0.160
2	0.054	1.195	0.482
3	0.032	-0.547	0.832
4	0.534	-5.952	0.038
5	0.034	0.218	0.808
6	0.007	0.085	0.924
7	0.051	0.499	0.665
8	0.107	0.795	0.461
9	0.132	-2.209	0.232
10	0.088	-0.653	0.592
11	0.032	0.210	0.816
12	0.136	-1.347	0.352
13	0.085	0.686	0.531
14	0.050	0.496	0.665
15	0.015	0.438	0.931
16	0.155	4.620	0.028
17	0.036	-0.295	0.847
18	0.122	0.671	0.482

表 4.6　安定冠動脈疾患患者に対する治療薬のベイズ流メタアナリシスの尺度混合正規分布における尺度パラメータの事後推定値

試験	二重指数分布		*t* 分布	
	事後平均 [95%信用区間]	$\Pr(\lambda_k \geqq 1)$	事後平均 [95%信用区間]	$\Pr(\lambda_k \geqq 1)$
1	2.54 [0.41, 7.53]	0.820	4.44 [0.47, 21.92]	0.785
2	1.67 [0.08, 6.23]	0.546	2.50 [0.26, 11.87]	0.518
3	1.32 [0.03, 5.64]	0.424	2.23 [0.23, 9.93]	0.454
4	5.22 [0.39, 12.40]	0.938	30.50 [1.08, 158.11]	0.977
5	1.85 [0.05, 6.98]	0.578	3.18 [0.26, 15.53]	0.578
6	1.89 [0.05, 6.99]	0.589	3.45 [0.27, 18.84]	0.595
7	1.50 [0.03, 6.00]	0.489	2.30 [0.24, 10.79]	0.487
8	1.80 [0.05, 6.58]	0.585	3.11 [0.26, 14.10]	0.568
9	2.38 [0.27, 7.26]	0.771	4.44 [0.40, 20.45]	0.749
10	1.80 [0.07, 6.57]	0.584	2.97 [0.29, 14.99]	0.604
11	1.80 [0.04, 6.78]	0.558	2.95 [0.26, 16.32]	0.574
12	2.22 [0.16, 7.12]	0.717	4.33 [0.36, 18.28]	0.719
13	1.68 [0.05, 6.31]	0.547	2.55 [0.24, 13.16]	0.537
14	1.49 [0.03, 6.00]	0.487	2.51 [0.23, 11.63]	0.489
15	1.10 [0.01, 5.33]	0.349	2.12 [0.22, 9.50]	0.431
16	3.35 [0.77, 8.78]	0.945	7.34 [0.73, 36.85]	0.932
17	1.43 [0.03, 5.78]	0.465	2.37 [0.23, 10.17]	0.477
18	2.07 [0.06, 7.38]	0.631	5.21 [0.27, 21.63]	0.637

4.3.3 妊娠糖尿病の既往の 2 型糖尿病発症リスク

この事例は，フォレストプロットから視覚的に外れ値がないと想定される事例であり(Vounzoulaki et al., 2020)，妊娠糖尿病患者の 2 型糖尿病の発症リスクを評価したメタアナリシスである．妊娠糖尿病は妊娠中に初めて診断される軽度の糖代謝異常であり，産後に 2 型糖尿病を発症するリスク因子として知られている．これまでの報告では人種差および追跡期間を考慮した発症リスクの定量評価が不十分であり，原著論文では正常血糖値の妊娠歴のある女性を対象とした 20 の研究により詳細な解析を行った．各研究における妊娠糖尿病および 2 型糖尿病は，各国の学会のガイドライン等による血糖値，ヘモグロビン A1c

（HbA1c），および経口糖負荷試験での血糖値の基準により診断された．対象とした 2 型糖尿病発症のリスク比を効果指標とした事例に対するベイズ流メタアナリシスの結果を図 4.4 に示す．試験間分散の事後平均は0.59であった．統合されたリスク比の事後平均は 9.92（95 %信用区間：6.54 ~ 15.39）であり，妊娠糖尿病の既往のある女性の発症リスクは正常血糖値の妊娠歴のある女性と比較して約 10 倍高いことが示唆された．

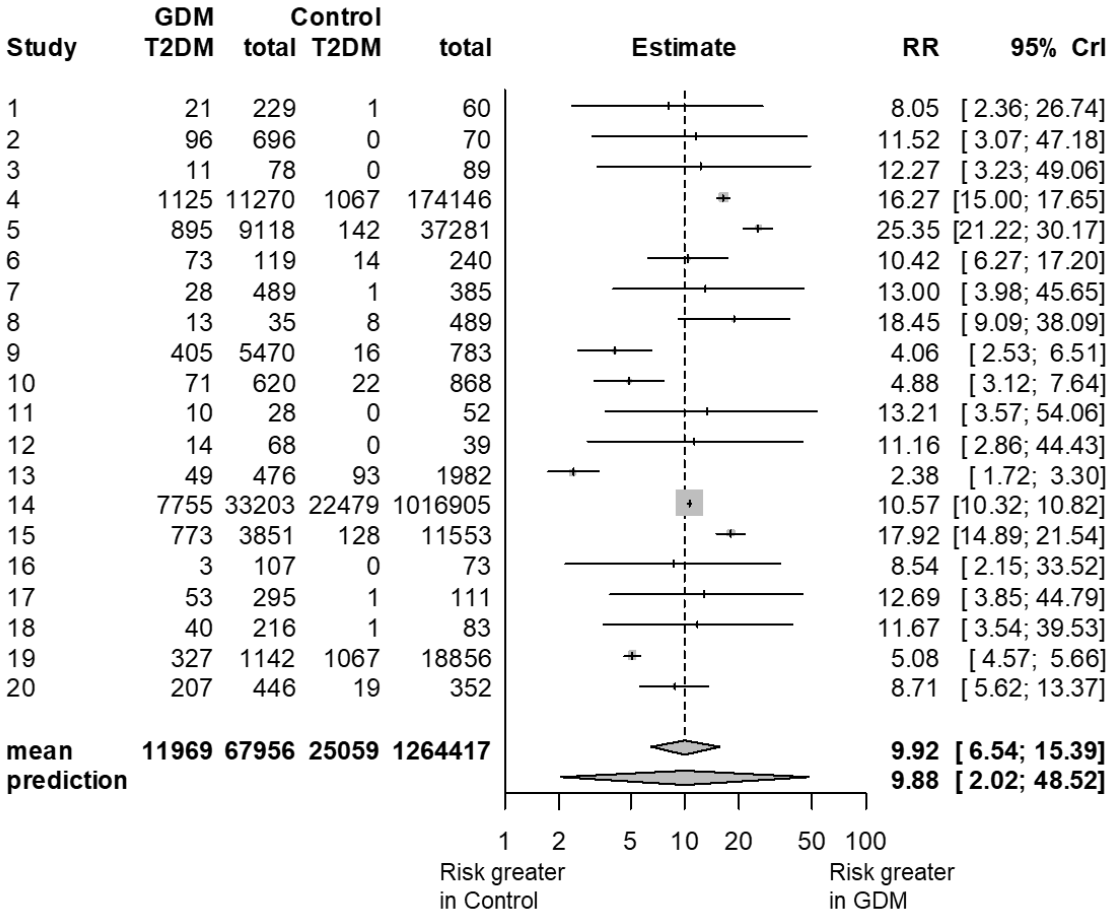

図 4.4　妊娠糖尿病の既往の 2 型糖尿病発症リスクを評価したベイズ流メタアナリシスにおける研究毎のリスク比についてのフォレストプロット

この事例の研究毎の相対距離，標準化残差，ベイズ流 P 値を表 4.7 に，尺度混合正規分布における尺度パラメータの事後推定値を表 4.8 に示す．相対距離はいずれの研究も低く，その中では研究 13 が 0.048 と最も高かった．標準化残差により外れ値として検出された試験は，研究 13 のみであった．また，ベイズ流 P 値および尺度混合正規分布における尺度パラメータの事後推定値により外れ値として検出された研究は，いずれも存在しなかった．

表 4.7　妊娠糖尿病の既往の 2 型糖尿病発症リスクを評価したベイズ流メタアナリシスにおける相対距離, 標準化残差, およびベイズ流 P 値

研究	相対距離	標準化残差	ベイズ流 P 値
1	0.009	-0.749	0.642
2	0.004	0.852	0.676
3	0.006	1.227	0.552
4	0.017	0.654	0.511
5	0.039	1.370	0.218
6	0.001	0.067	0.940
7	0.009	1.013	0.535
8	0.022	1.100	0.342
9	0.032	-1.383	0.221
10	0.027	-1.056	0.334
11	0.009	1.724	0.400
12	0.003	0.653	0.744
13	0.048	-2.380	0.066
14	0.000	0.073	0.928
15	0.021	0.805	0.430
16	0.006	-0.920	0.672
17	0.008	0.883	0.581
18	0.004	0.549	0.730
19	0.025	-0.904	0.384
20	0.007	-0.192	0.860

表 4.8　妊娠糖尿病の既往の 2 型糖尿病発症リスクを評価したベイズ流メタアナリシスの尺度混合正規分布における尺度パラメータの事後推定値

研究	二重指数分布		*t* 分布	
	事後平均 [95%信用区間]	Pr($\lambda_k \geqq 1$)	事後平均 [95%信用区間]	Pr($\lambda_k \geqq 1$)
1	1.59 [0.04, 6.46]	0.502	2.34 [0.25, 11.53]	0.523
2	1.42 [0.03, 5.74]	0.468	2.28 [0.24, 11.12]	0.487
3	1.52 [0.04, 5.86]	0.504	2.56 [0.25, 12.14]	0.511
4	2.01 [0.05, 7.34]	0.612	6.37 [0.26, 34.69]	0.621
5	2.02 [0.06, 7.40]	0.615	8.95 [0.27, 54.64]	0.645
6	1.86 [0.05, 6.80]	0.586	3.37 [0.26, 17.47]	0.585
7	1.62 [0.05, 6.31]	0.522	2.74 [0.25, 12.91]	0.531
8	1.93 [0.05, 6.82]	0.615	3.98 [0.27, 19.41]	0.617
9	2.11 [0.05, 7.45]	0.633	6.60 [0.28, 35.44]	0.671
10	2.05 [0.05, 7.37]	0.621	5.17 [0.28, 27.68]	0.647
11	1.76 [0.07, 6.52]	0.571	2.78 [0.27, 13.28]	0.559
12	1.39 [0.02, 5.65]	0.454	2.56 [0.23, 10.75]	0.492
13	2.40 [0.07, 8.65]	0.666	18.14 [0.30, 116.00]	0.734
14	2.00 [0.05, 7.36]	0.604	10.15 [0.27, 41.99]	0.633
15	1.96 [0.05, 7.33]	0.593	6.48 [0.27, 29.85]	0.627
16	1.49 [0.03, 6.10]	0.478	2.26 [0.24, 11.42]	0.497
17	1.55 [0.03, 5.98]	0.508	2.38 [0.25, 12.20]	0.523
18	1.47 [0.03, 5.83]	0.480	2.37 [0.24, 11.81]	0.497
19	2.06 [0.06, 7.54]	0.625	7.92 [0.27, 40.43]	0.637
20	1.86 [0.05, 6.87]	0.586	3.51 [0.26, 19.88]	0.602

4.3.4 新生児の呼吸窮迫症候群発症に対する出産前ステロイド投与

この事例は，新生児の呼吸窮迫症候群発症に対する妊婦への副腎皮質ステロイド投与の有効性を評価したメタアナリシスである(Saccone and Berghella, 2016)．新生児呼吸窮迫症候群は肺の発達が未熟であることによる呼吸障害であり，早産児によく見られる疾患である．副腎皮質ステロイド投与は胎児のサーファクタント産生を誘発するため，妊娠 24~33 週の妊婦への新生児呼吸窮迫症候群発症のリスクおよび重症度を低減することが報告されていたが，妊娠 34 週以降の妊婦における有効性は明確でなかった．原著論文では妊娠 34 週以上

で早産になりそうな単胎妊婦を対象としてプラセボまたは無治療を対照とした6のランダム化比較試験により詳細な解析を行った．対象とした事例に対するベイズ流メタアナリシスの結果を図4.5に示す．試験間分散の事後平均は0.56であった．統合されたリスク比の事後平均は0.73（95 %信用区間：0.38 ~ 1.52）であった．

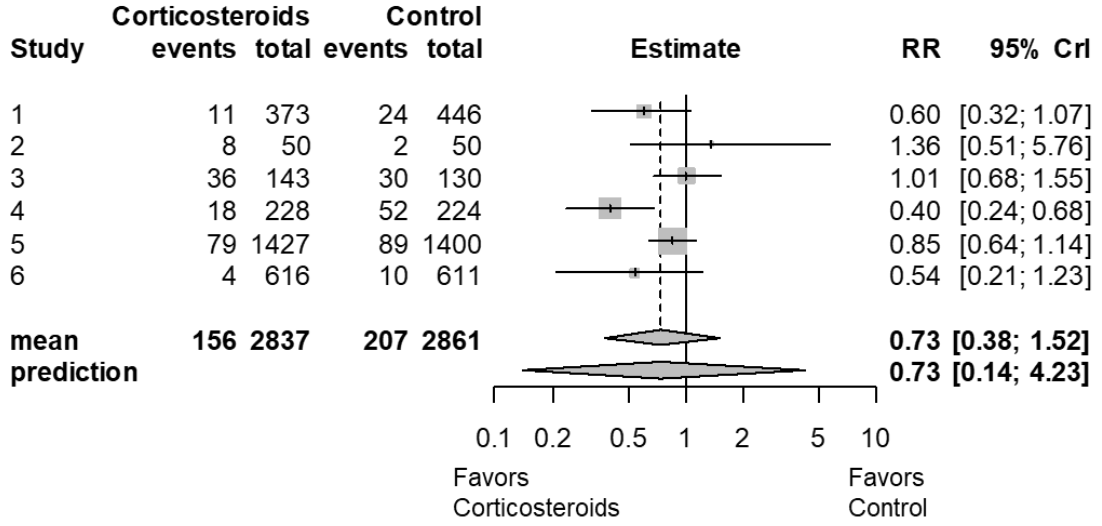

図4.5　新生児の呼吸窮迫症候群発症に対する出産前ステロイド投与のベイズ流メタアナリシスにおける試験毎のリスク比についてのフォレストプロット

この事例の試験毎の相対距離，標準化残差，ベイズ流P値を表4.9に，尺度混合正規分布における尺度パラメータの事後推定値を表4.10に示す．相対距離は試験4が0.596と最も高かった．標準化残差により外れ値として検出された試験は，試験2のみであった．ベイズ流P値および尺度混合正規分布における尺度パラメータの事後推定値により外れ値として検出された試験は，いずれも存在しなかった．

表 4.9　新生児の呼吸窮迫症候群発症に対する出産前ステロイド投与のベイズ流メタアナリシスにおける相対距離, 標準化残差, およびベイズ流 P 値

試験	相対距離	標準化残差	ベイズ流 P 値
1	0.233	-0.317	0.683
2	0.462	2.520	0.118
3	0.306	0.478	0.535
4	0.596	-1.354	0.275
5	0.106	0.184	0.765
6	0.293	-0.688	0.491

表 4.10　新生児の呼吸窮迫症候群発症に対する出産前ステロイド投与の有効性を評価したベイズ流メタアナリシスの尺度混合正規分布における尺度パラメータの事後推定値

試験	二重指数分布		*t* 分布	
	事後平均 [95%信用区間]	$\Pr(\lambda_k \geqq 1)$	事後平均 [95%信用区間]	$\Pr(\lambda_k \geqq 1)$
1	1.74 [0.05, 6.68]	0.558	3.13 [0.25, 14.78]	0.569
2	2.82 [0.24, 8.31]	0.830	5.73 [0.40, 30.03]	0.821
3	2.10 [0.06, 7.36]	0.638	4.47 [0.28, 24.24]	0.644
4	2.39 [0.07, 8.06]	0.695	6.23 [0.30, 33.33]	0.722
5	1.98 [0.05, 7.28]	0.609	4.21 [0.27, 23.81]	0.619
6	1.78 [0.05, 6.64]	0.575	3.38 [0.27, 14.77]	0.577

4.3.5 感度分析

これまでの事例に関して各指標で外れ値として検出された試験を除外した感度分析の結果を表 4.11 に示す．4.3.1 項の事例では，効果指標が平均差である場合，二重指数分布の尺度パラメータの事後推定値以外で共通して外れ値と判定された試験 16 を除外することで平均治療効果が増加し，試験間分散が大幅に減少した．また，効果指標が標準化平均差である場合，標準化残差およびベイズ流 P 値で外れ値として検出された試験 16 および試験 18 を除外することで平均治療効果は大幅に増加し，試験間分散は大幅に減少した．試験 23 をさらに

除外することによる平均治療効果および試験間分散に対する影響はわずかであった．4.3.2 項の事例では，二重指数分布の尺度パラメータの事後推定値以外で共通して外れ値と判定された試験 4 を除外することで平均治療効果が増加し，試験間分散が減少した．4.3.3 項および 4.3.4 項の事例では，いずれも標準化残差により外れ値と判定された 1 試験を除外したが，平均治療効果および試験間分散に対する影響はわずかであった．

表 4.11　外れ値として検出された試験を除外した感度分析結果

事例	効果指標	指標	外れ値として検出された試験	平均治療効果 事後平均[95%信用区間]	試験間分散 事後平均[95%信用区間]
4.3.1 項	平均差	全試験	−	-3.18 [-7.97, 1.61]	114.65 [55.21, 222.00]
		標準化残差	16, 23	-1.10 [-4.23, 2.22]	38.79 [16.80, 81.18]
		ベイズ流 P 値	16	-1.62 [-4.91, 1.73]	44.34 [18.25, 94.66]
		尺度パラメータ（二重指数分布）	−	−	−
		尺度パラメータ（*t* 分布）	16, 18, 23	-0.53 [-2.47, 1.81]	7.83 [0.04, 26.52]
	標準化平均差	全試験	−	-0.27 [-0.67, 0.10]	0.80 [0.36, 1.61]
		標準化残差	16, 18	-0.03 [-0.15, 0.09]	0.04 [0.01, 0.11]
		ベイズ流 P 値	16, 18	-0.03 [-0.15, 0.09]	0.04 [0.01, 0.11]
		尺度パラメータ（二重指数分布）	16, 18, 23	-0.01 [-0.13, 0.11]	0.03 [0.01, 0.09]
		尺度パラメータ（*t* 分布）	16, 18, 23	-0.01 [-0.13, 0.11]	0.03 [0.01, 0.09]
4.3.2 項	死亡率比	全試験	−	0.82 [0.65, 1.03]	0.13 [0.03, 0.37]
		標準化残差	1, 4, 9, 16	0.92 [0.81, 1.04]	0.02 [0.00, 0.07]
		ベイズ流 P 値	4, 16	0.90 [0.79, 1.02]	0.02 [0.00, 0.08]
		尺度パラメータ（二重指数分布）	−	−	−
		尺度パラメータ（*t* 分布）	4	0.91 [0.80, 1.04]	0.02 [0.00, 0.09]
4.3.3 項	リスク比	全試験	−	9.92 [6.54, 15.39]	0.59 [0.25, 1.29]
		標準化残差	13	11.08 [7.59, 16.38]	0.42 [0.16, 1.00]
		ベイズ流 P 値	−	−	−
		尺度パラメータ（二重指数分布）	−	−	−
		尺度パラメータ（*t* 分布）	−	−	−
4.3.4 項	リスク比	全試験	−	0.73 [0.38, 1.52]	0.56 [0.03, 2.78]
		標準化残差	2	0.63 [0.32, 1.15]	0.41 [0.02, 1.99]
		ベイズ流 P 値	−	−	−
		尺度パラメータ（二重指数分布）	−	−	−
		尺度パラメータ（*t* 分布）	−	−	−

4.4 考察

メタアナリシスでは外れ値となる試験が混入する可能性があり，外れ値の存在下において推奨されている感度分析を行うため，どの試験が外れ値であるか，またその影響力を診断できることが求められる．また，治療法のエビデンスを構築するためのメタアナリシスとして，近年は複数の治療法を同時に評価できるネットワークメタアナリシスにおける手法も開発されてきているものの (Noma et al., 2020; Zhang et al., 2015)，最も広く用いられているのは単変量メタアナリシスである．本章で提案した手法は，ベイズ流の解析をする際，単変量メタアナリシスにおける外れ値問題に対して外れ値の検出およびその影響力の診断により，感度分析で除外する対象とすべき試験を提示することができると考える．

相対距離，標準化残差，ベイズ流 P 値，および尺度混合正規分布における尺度パラメータの事後推定値の特性を 4 つの事例を通じて確認した．相対距離はメタアナリシスにおいて最も関心のある平均治療効果に対する影響力を反映するが，試験が外れ値であるかどうかを相対距離単独で判断することはできない．一方，標準化残差は試験の治療効果の推定値と新たな試験の治療効果の事後予測分布における平均との距離を反映する指標であり，事例解析においては視覚的に外れていると判断される試験を検出することができた．しかし，4.3.2 項および 4.3.3 項の事例のように相対距離が 0.1 程度の試験も検出されており，平均治療効果に対する影響力の低い試験であっても検出された．また，ベイズ流 P 値は事後予測モデルチェックとして事前分布や尤度により仮定したモデルにおける試験の治療効果の推定値の極端さを反映する指標である．事例解析においては標準化残差より検出された試験は少ないものの，視覚的に外れていて相対距離が 0.4 以上の平均治療効果に対する影響力の高い試験は検出すること

ができた．そして，尺度混合正規分布は試験毎の治療効果の指標の分布のモデルに極端な値をとることを許容したモデルであり，主要な解析に用いられる正規-正規モデルの仮定に対する影響力を評価する手法である．Carlin and Louis (2009) によって提案されてからベイズ流の影響力診断の手法として長らく多くの応用問題に使われており，事例解析においては 4.3.1 項で効果指標を平均差とした事例で治療効果の推定値が平均治療効果の推定値から離れており相対距離が 0.265 と影響力の比較的高い試験 18 が，標準化残差およびベイズ流 P 値で外れ値として検出されず，*t* 分布を仮定した尺度パラメータの事後推定値で検出されていた．この理由として，試験 18 がこの事例で最も外れている試験でないことで標準化残差およびベイズ流 P 値では外れ値として検出できなかったが，*t* 分布を仮定した尺度パラメータの事後推定値では分布の形状を考慮していることで検出できたためと考えられる．また，使用した分布として，*t* 分布の方が二重指数分布より裾が重く外れ値を加味できる分布であり，今回の事例解析では外れ値として検出された試験数が多かった．試験数が少ない場合の指標の特性として試験数が 6 である 4.3.4 項の事例では，標準化残差のみ試験 2 を外れ値として検出しており，試験 2 の治療効果の推定値は平均治療効果の事後平均から最も離れている試験であった．標準化残差は交差検証に基づく手法であるのに対し，ベイズ流 P 値および尺度混合正規分布における尺度パラメータの事後推定値は検証する試験を含むモデルを構築しているため，保守的な結果が得られることが想定される．試験数が少ない事例においては，ベイズ流 P 値および尺度混合正規分布における尺度パラメータの事後推定値が，外れ値を検出できない可能性があるため注意が必要である．また，これら 4 種類の指標が外れ値をどの程度正確に検出できるかを検討していないため，今後，シミュレーション実験等による動作特性の検証が必要である．

指標により外れ値として検出された試験を除外した感度分析では，いずれの事例においても平均治療効果の事後分布による結論が変わることはなかったものの，多くの指標で外れ値として検出された試験を除外することで平均治療効果は大きく変化した．ただし，外れ値を検出する手法を用いる際，検出された外れ値を解析対象からそのまま除外することはできないことに注意する必要がある．このような手法は多くの解析を行うため，偶然にも外れ値として検出されてしまう可能性がある．しかし，外れ値を検出する絶対的な指標は存在しないので，回帰分析における手法と同様に本章で提案した複数の指標を用いていずれかの指標で外れ値として検出された試験を外れ値の可能性がある試験として調査することが望ましいと考えられる．なお，実際に除外すべきでない試験が外れ値として検出されても以後の調査の結果で除外するかどうかを判断するため，多重性の調整は不要であると考えられる．本章の事例解析では，外れ値と検出された試験が実際に除外されるべき試験であったかは確認できなかったが，このような指標により外れ値が検出された場合，その試験の原著論文を調査したり著者に問い合わせるといった対応が求められる．また，外れ値の原因を探索している過程で，新たな知見が得られることや，解析において加えるべき共変量が見つかることで研究が発展する可能性もあるため，指標により外れ値を客観的に検出することには意義があると考えられる．

メタアナリシスで用いられる手法は実践する非統計家の研究者にとって利用しやすいことが求められるが，本章で提案した指標は容易に計算することができ，外れ値および影響力の大きい試験を判定しやすいと考えられる．このような指標がメタアナリシスの実践の場で普及することで研究者が外れ値を慎重に取り扱うことができ，その結果，医療における適切な意思決定に貢献できると考える．

第 5 章 結論

情報通信技術の飛躍的な進歩により，世界中の最新の科学的知見の迅速な共有が可能となり，システマティックレビューによる医学的エビデンスの詳細な分析は，今後，ますます重要性を増すものと思われる．また，その方法論は，医学のみならず，社会学・教育学等でも広く用いられるようになっており，さまざまな社会的問題の解決に資するための技術評価の手法となることが予想される．

本研究では，メタアナリシスにおけるベイズ流の方法において，2 つの重要な研究課題を取り上げ，それを解決するための研究を行った．ベイズ流の予測区間は，本研究で扱った従来の対比較のメタアナリシスだけではなく，診断法の性能を評価するための 2 変量メタアナリシス(Deeks, 2001; Leeflang et al., 2008)や，複数の治療法の有効性を比較評価するためのネットワークメタアナリシス(Salanti, 2012)においても，広く用いられており，これらの研究手法においても，同様の議論が必要であると思われる．ただし，本研究から得られた結論は，概ね，多変量モデルを用いるこれらの方法にも一般化することができ，同様に成り立つことが予想される．詳細な分析については，今後の研究課題となると考えられる．いずれにしても，本研究で示したエビデンスは，ベイズ流メタアナリシスから得られる予測区間の正しい解釈を行うために，重要な指針を与えるものと考えられる．

また，外れ値の検出と影響力解析の方法については，ベイズ流メタアナリシスにおいて，Carlin-Louis 式の明確な評価方法の枠組みを与えたものとなっており，従来，適切な評価方法が存在しなかった問題に，新たに有用な手法を提示することができたものと考えられる．潜在的な外れ値の問題は，エビデンスの不正確性・不確実性にも関わる問題であり(Matsushima et al., 2020; Noma et al.,

2020)，本研究で開発した方法によって，システマティックレビューの実務，および，そのエビデンスの解釈を行う上で，これまでにない重要な情報を与えることができるものになるとも考えられる．一方で，これらの方法は，MCMC を用いた解析方法となっており，非統計家にも扱いやすい計算ツールの開発や，その普及のための事例研究や情報発信等も，今後の課題として重要であると考える．

本研究から得られた成果は，今後，ますます標準的な方法として普及することが予想されるベイズ流メタアナリシスの実践において，従来にない有益な知見をもたらすものであると考えられ，医療技術評価や診療ガイドラインの作成，医療政策の策定等，さまざまな重要な科学的評価・意思決定の問題に資するものであると考えられる．

謝辞

博士課程に関して，主任指導教員を引き受けて頂き，懇切丁寧にご指導頂きました統計数理研究所の野間久史 准教授に深く御礼申し上げます.

学位審査に関して，ご多忙の中，審査を引き受けて頂きました統計数理研究所の二宮嘉行 教授，逸見昌之 准教授，国立感染症研究所の米岡大輔 先生，筑波大学の丸尾和司 准教授に心より感謝申し上げます．逸見昌之 准教授には副指導教員もご担当頂き，重ねて感謝申し上げます.

本研究の投稿論文に関して，共著者としてご指導頂きました京都大学の古川壽亮 教授，King's College London および東京大学の山田朋英 先生，慶應義塾大学病院の長島健悟 先生，大塚製薬株式会社の松嶋優貴 先生に，厚く御礼申し上げます.

最後に，応援して頂き，支えて頂いた家族に心より感謝致します.

参考文献

Agresti, A., and Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 contingency tables. *Biometrics* **61**, 515-523.

Baker, R., and Jackson, D. (2008). A new approach to outliers in meta-analysis. *Health Care Management Science* **11**, 121-131.

Baker, R., and Jackson, D. (2016). New models for describing outliers in meta-analysis. *Research Synthesis Methods* **7**, 314-328.

Bangalore, S., Fakheri, R., Wandel, S., Toklu, B., Wandel, J., and Messerli, F. H. (2017). Renin angiotensin system inhibitors for patients with stable coronary artery disease without heart failure: systematic review and meta-analysis of randomized trials. *BMJ* **356**, j4.

Beath, K. J. (2014). A finite mixture method for outlier detection and robustness in meta-analysis. *Research Synthesis Methods* **5**, 285-293.

Berger, J. O., and Deely, J. (1988). A bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *Journal of the American Statistical Association* **83**, 364-373.

Berger, J. O., and Pericchi, L. R. (2001). *Objective Bayesian Methods for Model Selection: Introduction and Comparison*, In P. Lahiri edition. Beachwood, OH: Institute of Mathematical Statistics.

Bodnar, O., Link, A., Arendacká, B., Possolo, A., and Elster, C. (2017). Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine* **36**, 378-399.

Brockwell, S. E., and Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20**, 825-840.

Brockwell, S. E., and Gordon, I. R. (2007). A simple method for inference on an overall effect in meta-analysis. *Statistics in Medicine* **26**, 4531-4543.

Carlin, B. P., and Louis, T. A. (2009). *Bayesian Methods for Data Analysis.* New York, NY: Chapman & Hall.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* **19**, 15-18.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2009). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd edition. New York, NY: Russell Sage Foundation.

Daniels, M. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics* **27**, 567-578.

Deeks, J. J. (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* **323**, 157-162.

DerSimonian, R., and Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177-188.

DuMouchel, W. H., and Normand, S. L. (2000). *Computer Modeling Strategies for Meta-Analysis*, In D. K. Stangl & D. A. Berry edition. Boca Raton, FL: CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd edition. Boca Raton, FL: Chapman and Hall/CRC.

Gumedze, F. N., and Jackson, D. (2011). A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Medical Research Methodology* **11**, 19.

Häuser, W., Bernardy, K., Uçeyler, N., and Sommer, C. (2009). Treatment of fibromyalgia syndrome with antidepressants a meta-analysis. *JAMA* **301**, 198-209.

Higgins, J. P., and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539-1558.

Higgins, J. P. T., Thomas, J., Chandler, J.*, et al.* (2019). *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd edition. Chichester, UK: Wiley-Blackwell.

Higgins, J. P. T., Thompson, S. G., and Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A* **172**, 137-159.

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., and Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* **6**, e010247.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* **4**, 227-241.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London Series A* **186**, 453-461.

Julious, S. A., and Whitehead, A. (2012). Investigating the assumption of homogeneity of treatment effects in clinical studies with application to meta-analysis. *Pharmaceutical Statistics* **11**, 49-56.

Kontopantelis, E., Springate, D. A., and Reeves, D. (2013). A re-analysis of the Cochrane library data: The dangers of unobserved heterogeneity in meta-analyses. *PloS One* **8**, e69930.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* **24**, 2401-2428.

Lee, K. J., and Thompson, S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine* **27**, 418-434.

Leeflang, M. M., Deeks, J. J., Gatsonis, C., and Bossuyt, P. M. (2008). Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine* **149**, 889-897.

Lin, L. (2019). Use of prediction intervals in network meta-analysis. *JAMA Network Open* **2**, e199735.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* **28**, 3049-3067.

Matsushima, Y., Noma, H., Yamada, T., and Furukawa, T. A. (2020). Influence diagnostics and outlier detection for meta-analysis of diagnostic test accuracy. *Research Synthesis Methods* **11**, 237-247.

Nagashima, K., Noma, H., and Furukawa, T. A. (2019). Prediction intervals for random-effects meta-analysis: A confidence distribution approach. *Statistical Methods in Medical Research* **28**, 1689-1702.

Noma, H. (2011). Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Statistics in Medicine* **30**, 3304-3312.

Noma, H., Gosho, M., Ishii, R., Oba, K., and Furukawa, T. A. (2020). Outlier detection and influence diagnostics in network meta-analysis. *Research Synthesis Methods* **11**, 891-902.

Partlett, C., and Riley, R. D. (2017). Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Statistics in Medicine* **36**, 301-317.

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., and Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* **58**, 982-990.

Riley, R. D., Higgins, J. P. T., and Jonathan, J. D. (2011). Interpretation of random effects meta-analyses. *BMJ* **342**, d549.

Röver, C. (2020). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software* **93**, 1-51.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151-1172.

Rubinstein, S. M., de Zoete, A., van Middelkoop, M., Assendelft, W. J. J., de Boer, M. R., and van Tulder, M. W. (2019). Benefits and harms of spinal manipulative therapy for the treatment of chronic low back pain: systematic review and meta-analysis of randomised controlled trials. *BMJ* **364**, l689.

Saccone, G., and Berghella, V. (2016). Antenatal corticosteroids for maturity of term or near term fetuses: systematic review and meta-analysis of randomized controlled trials. *BMJ* **355**, i5044.

Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods* **3**, 80-97.

Salvo, F., Moore, N., Arnaud, M.*, et al.* (2016). Addition of dipeptidyl peptidase-4 inhibitors to sulphonylureas and risk of hypoglycaemia: systematic review and meta-analysis. *BMJ* **353**, i2231.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14**, 2685-2699.

Snell, K. I. E., Hua, H., Debray, T. P. A.*, et al.* (2016). Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *Journal of Clinical Epidemiology* **69**, 40-50.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* . Chichester, UK: John Wiley & Sons.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604-608.

Veroniki, A. A., Jackson, D., Bender, R.*, et al.* (2019). Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Research Synthesis Methods* **10**, 23-43.

Viechtbauer, W., and Cheung, M. W. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* **1**, 112-125.

Vounzoulaki, E., Khunti, K., Abner, S. C., Tan, B. K., Davies, M. J., and Gillies, C. L. (2020). Progression to type 2 diabetes in women with a known history of gestational diabetes: systematic review and meta-analysis. *BMJ* **369**, m1361.

Wang, M. C., and Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods* **3**, 46-54.

Whitehead, A. (2002). *Meta-Analysis of Controlled Clinical Trials*. Chichester, UK: Wiley.

Whitehead, A., and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* **10**, 1665-1677.

Zhang, J., Fu, H., and Carlin, B. P. (2015). Detecting outlying trials in network meta-analysis. *Statistics in Medicine* **34**, 2695-2707.