博士論文

国際共同治験における外れ値となる地域の検出と影響力診断の方法

2022 年 3 月

総合研究大学院大学
複合科学研究科統計科学専攻

青木 誠

概要

医薬品の開発においては，新薬開発の効率化・迅速化のため，複数の地域で同時に一つの臨床試験を実施する国際共同治験が増加してきている。国際共同治験の主な目的は地域共通の治療効果を検証する事であるが，新薬の治療効果に関連する地域特有の遺伝子多型や医療習慣などの要因により，地域間で治療効果が異なる場合がある。そのため，地域共通の治療効果において有効性が示せた場合に各地域でも同様の治療効果があるかどうかの一貫性を評価し，一貫していない場合にその影響を評価する事が国際共同治験を実施する上で重要な統計的課題となっている。そこで，本研究では回帰分析やメタアナリシスにおいてこれまで研究されてきた外れ値の検出とその影響力を評価する方法を国際共同治験の一貫性評価の枠組みに応用し，（1）Leave-one-out cross-validation 型のスチューデント化残差に基づく方法，（2）尤度比検定に基づく方法，（3）全体分散の相対的変化に基づく方法，（4）地域間分散の相対的変化に基づく方法を提案する。また，これらの提案法の統計量に対して，基準を設定し，統計的有意性を評価する際に，統計量のばらつきを考慮するため，帰無仮説の下での統計量の分布の推定にパラメトリックブートストラップ法を適用する。さらに，リバーロキサバン，ロサルタン，メトプロロールの 3 つの薬剤に対してそれぞれ実施された国際共同治験の事例に対して 4 つの提案法を応用し，実践的な有用性を検証する。最後に，事例を基にシナリオを設定し，シミュレーションにより提案法の性能を評価する。

目次

1. 序論

1.1 各国の医薬品開発における歴史的背景

医薬品の開発においては，1998 年頃までは各地域（国）でそれぞれ臨床試験を行い，その臨床試験データに基づいて新薬の有効性及び安全性を示す事により，新薬の承認が取得されてきた。この時，人口の多い地域では患者数も多く，臨床試験で容易に必要な患者数を集める事が可能であり，早期に臨床試験を終え，新薬が承認されてきた。そのため，国間で承認時期に差が生じ，開発の進んでいる地域で使用できる新薬が別の地域では使用できないという問題（ドラッグ・ラグ）が指摘されてきた。そこで，1998 年に日米 EU 医薬品規制調和国際会議（以下，ICH）より「外国臨床データを受け入れる際に考慮すべき民族的要因についての指針」に関するガイドラインが発出された（ICH 1998）。このガイドラインの発出により，単一地域のみで十分な患者数を集める事が困難な地域では，他地域の臨床試験データを外挿する事により臨床データの国際的な重複を最小限にし，患者へ有益な医薬品を迅速に提供する事が可能となった。しかしながら，用量反応性や安全性，有効性に関しては自国のデータが他地域のデータと類似している事を示す必要があり，自国でも臨床試験を実施する必要があったため，依然としてドラッグ・ラグの課題が解消されずにいた（Shimatani and Sudo 2005）。そこで，2007 年頃から更なる開発の効率化・迅速化のため，複数の地域で同時に一つの臨床試験を実施する国際共同治験が増加してきた（Asano et al. 2013, Ichimaru et al. 2010）。さらに，2017 年には ICH より「国際共同治験の計画及びデザインに関する一般原則」に関するガイドラインが発出された（ICH 2017）。ここでは主に世界各地域での承認申請において国際共同治験の受け入れ可能性を高めるために，国際共同治験の計画及びデザインの一般原則が示されており，今後も更なる国際共同治験の増加が見込まれる。

1.2 国際共同治験の課題

国際共同治験では多地域で共通の治験実施計画書を用いて全ての参加地域で共通の新薬の治療効果を示す事が主目的となっている。（ICH 2017, Quan et al. 2013, Quan et al. 2017）。しかしながら，地域間では様々な新薬の治療効果に関連する民族的要因が不均一であるため，異なる地域での治療効果は一貫していない可能性がある（ICH 1998, Tohkin 2016）。また，地域ごとの被験者数の違いや民族的要因による治療効果の違いによって，全体の結果を歪めてしまうような影響力のある地域が存在する可能性がある。そのため，全体集団で有意な治療効果が示されたとしても，地域間で一貫した結果が得られなかったり，極端に影響力のある地域が存在する場合，それらの地域で治験薬が有効でない又は高い有効性がある可能性を疑うべきである。なお，ここでの治療効果に影響を与える民族的要因は臨床試験開始前に特定されている事が理想的であるが，図 1 に示す通り，多くの内因性と外因性の側面を考慮する必要があり，特定は容易ではない。そのため，ここでは内因性及び外因性を代表する「地域」を単位として考える。一貫した結果が得られなかった地域や影響力のある地域は，全ての参加地域共通の治療効果の推定や臨床的解釈，規制当局の承認審査における意思決定に大きな影響を及ぼす可能性がある。したがって，地域間の治療効果の一貫性を評価する

事は国際共同治験の重要な統計的課題となっている（Chen et al. 2010, Chen et al. 2011, Diao et al. 2017, Guo et al. 2016, Liu et al. 2016, Quan et al. 2010b, Quan et al. 2013, Quan et al. 2014, Teng et al. 2018, Tsou et al. 2012）。そこで，これまで地域間の治療効果の一貫性を十分な検出力で評価するための各参加地域への被験者数の割り当て方法について議論されている（Diao et al. 2017, Ikeda and Bretz 2010, Kawai et al. 2008, Ko et al. 2010, Quan et al. 2010b, Quan et al. 2013, Quan et al. 2014, Uesaka 2009）。日本においては厚生労働省より「国際共同治験に関する基本的考え方について」の通知（厚生労働省 2007）が発出されており，地域間の一貫性を評価する方法及びそれに基づく被験者数の割り当て方法について言及されている。この方法においては各地域で十分な被験者数が要求されるが，複数の地域で競合的に被験者登録が行われる状況では十分な被験者数を確保する事は容易ではない。また，探索的に民族差を検討する上では被験者規模に縛られない，より柔軟な方法が必要である。さらに，この方法では各地域の治療効果の推定値が基準値を超えたかどうかの定性的な評価を与えるだけであり，一貫していない地域の地域共通の治療効果への定量的な影響を評価する事は困難である。そこで，本研究では一貫しない地域を評価するため，全体と異なる治療効果を持つ地域を潜在的な外れ値となる地域と考え，回帰分析やメタアナリシスにおいてこれまで研究されている外れ値の検出と影響力の評価の方法の応用を考える。これまで数理統計学や応用統計学で回帰分析において外れ値を検出し，それらの外れ値が回帰モデルにどのように影響するかを判断するための効果的な方法が研究されている（Belsley et al. 1980）。また，メタアナリシスの枠組みにおいても，外れ値となる研究の検出や影響力のある研究の特定に関して研究が進んでいる（Negeri and Beyene 2020, Noma et al. 2020, Viechtbauer and Cheung 2010）。一貫しない地域を外れ値となる地域として検出する方法はこれまで提案されていない。本論文では，2 章で国際共同試験における従来の各地域の治療効果の評価方法を説明し，3 章で外れ値となる地域及び影響力のある地域の検出法について紹介する。最後に 4 章で結論を述べる。

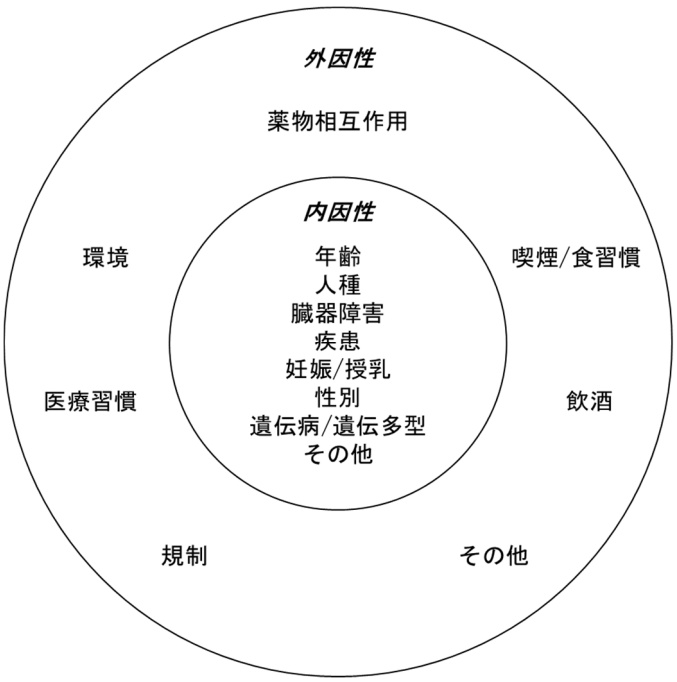

図 1 医薬品の曝露量や反応に影響を及ぼす民族的要因（小山，山本 2016）

2. 従来の各地域の治療効果の評価方法

国際共同治験における主要な目的は全地域共通の治療効果を検証する事である。通常，全地域共通の治療効果の有効性が示せた場合，各地域で同様の治療効果があるかどうかの一貫性の評価や治療効果と地域の交互作用を確認する。この治療効果の一貫性の評価や地域の交互作用の評価においては，各地域の要約統計量を用いて評価を行うか，すべての参加地域の個々の被験者データに基づき，共分散分析，ロジスティック回帰モデル，コックス比例ハザードモデルなどを用いて評価を行う（Uesaka 2009）。そこで，最近の理論的研究で，メタアナリシスにおける各研究の治療効果の推定値（要約統計量）を併合した治療効果の最尤推定量が個々の被験者データから推定した治療効果の最尤推定量と同様の漸近効率をもつ事が明らかになっている事より（Lin and Zeng 2010），解析をより単純にするため，メタアナリシスにおける「研究」を国際共同治験の「地域」に置き換えて，各地域の治療効果の要約統計量を用いて評価する事とする。この時，$y_i(i = 1, \cdots, k)$をi番目の地域の治療効果の推定値（例えば，平均値の差，対数オッズ比，対数ハザード比など）とする。

2.1 固定効果モデルと変量効果モデル

通常，全地域共通の治療効果の推定には固定効果モデルや変量効果モデルが使われている（Kim and Kang 2019）。固定効果モデルにおいては，各地域の治療効果を$\theta_i(i = 1, \cdots, k)$とした時，$\theta_1 = \theta_2 = \cdots = \theta_k = \theta$を仮定し，$\theta$は共通の治療効果パラメータとする。この時，大標本近似を用いてy_iに以下の正規分布を仮定できる（Quan et al. 2013）。

$$y_i \sim N(\theta, \sigma_i^2) \tag{1}$$

ここで，σ_i^2は部分集団解析から得られた各地域の分散とし，この分散により各地域の分散は既知であると仮定する。

また，変量効果モデルにおいては，各地域の治療効果θ_iが異なる事を仮定し，以下の通り，θ_iに正規分布を仮定できる（Chen et al. 2012, Hung et al. 2010, Kim and Kang 2019, Liu et al. 2016, Quan et al. 2010a）。

$$y_i \sim N(\theta_i, \sigma_i^2), \theta_i \sim N(\mu, \tau^2) \tag{2}$$

ここで，μはk個の地域の治療効果の全体平均，τ^2は地域間分散とする。変量効果モデルにおいては各地域のデータに基づき異質性を表現する事ができる。モデルパラメータの推定に関してはメタアナリシスの分野で様々な有効な推定法が提案されており（Veroniki et al. 2016, Viechtbauer 2005），国際共同治験において各地域のモデルパラメータを推定する場合でも同様の推定法が適用できる。本論文では，標準的な方法として固定効果モデルには逆分散法を採用し，変量効果モデルには制限付き最尤法（REML）を採用したが，他の推定法も適用可能である。なお，本章にていくつかのモデルパラメータの推定法を示す。固定効果モデル(1)は変量効果モデル(2)の$\tau^2 = 0$の場合に対応するため，3 章以降では主に変量効果モデル(2)に基づく手法を説明する。

2.1.1. 逆分散法

メタアナリシスで一般に用いられる併合した治療効果を推定する単純な方法として，逆分散法がある。この方法は離散データでも連続データでも同様に適用する事が可能である。逆分散法では各研究の治療効果を併合する際の重みを各研究における治療効果の推定値の分散の逆数としている。そのため，より被験者数の大きい研究はより小さな標準誤差を持ち，より大きな重みを得る事となる。この重みを用いる事で併合した治療効果の不確実性を最小化する事ができる。この時，固定効果モデルの場合，各研究の治療効果の重み$w_i(i = 1, \cdots, k)$は，

$$w_i = \frac{1}{\sigma_i^2} \tag{3}$$

となり，統合した治療効果及びその分散は，

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \tag{4}$$

$$Var(\hat{\theta}) = \frac{1}{\sum_{i=1}^{k} w_i} \tag{5}$$

で推定できる。変量効果モデルの場合，各研究の治療効果の重みw_iは，

$$w_i^* = \frac{1}{\sigma_i^2 + \hat{\tau}^2} \tag{6}$$

となる。この時，$\hat{\tau}^2$の推定法については後述する DerSimonian-Laird 推定量などを適用する事ができる。統合した治療効果及びその分散は，

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i^* y_i}{\sum_{i=1}^{k} w_i^*} \tag{7}$$

$$Var(\hat{\mu}) = \frac{1}{\sum_{i=1}^{k} w_i{}^*} \tag{8}$$

で推定できる（Borenstein et al. 2010, DerSimonian and Laird 1986, Higgins et al. 2019）。

2.1.2. DerSimonian-Laird 推定量

DerSimonian-Laird 推定量は反復計算がなく，単純に計算できるため，最も頻繁にτ^2の推定に用いられるアプローチである。Cochran の Q 統計量が

$$Q = \sum_{i=1}^{k} w_i \left(y_i - \hat{\theta}\right)^2 = \sum_{i=1}^{k} \frac{\left(y_i - \hat{\theta}\right)^2}{\sigma_i^2} \tag{9}$$

で表され，DirSimonian-Laird 推定量は

$$\hat{\tau}^2 = \max\left(0, \frac{Q - (k-1)}{\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 \,/\, \sum_{i=1}^{k} w_i}\right) \tag{10}$$

で推定される（DerSimonian and Laird 1986）。$\theta_1 = \theta_2 = \cdots = \theta_k$の仮定の下では Cochran の Q 統計量は自由度$k-1$のχ^2分布に従うため，Cochran の Q 統計量の期待値は$k-1$となり，$\hat{\tau}^2 = 0$となる。

2.1.3. 最尤（ML）推定量

ML 推定量は漸近的に有効だが，反復計算が必要となる。変量効果モデル(2)に基づいて，対数尤度関数は以下の式で表される（Veroniki et al. 2016）。

$$l(\mu,\tau^2) = -\frac{1}{2}\sum_{i=1}^{k}\left\{\log 2\pi(\sigma_i^2+\tau^2) + \frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2}\right\} \tag{11}$$

μ，τ^2の ML 推定量は$l(\mu,\tau^2)$を最大化する事により得られる。これはμとτ^2で偏微分を行い，その式が 0 となるようなμとτ^2が ML 推定量となる。最大化の方法としては初期値を与え，反復計算により ML 推定量が収束するまで繰り返す。$\hat{\tau}^2$の初期値としては，反復計算を必要としない方法に基づく推定量を設定するか，$\hat{\tau}^2 = 0$などを設定できる。

2.1.4. 制限付き最尤（REML）推定量

REML 推定量で用いる変量効果モデル(2)に基づいて，制限付き対数尤度関数は以下のように記述される（Veroniki et al. 2016）。

$$l_{RL}(\mu,\tau^2) = -\frac{1}{2}\left\{\sum_{i=1}^{k}\left[\frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2} + \log\left(\sigma_i^2+\tau^2\right)\right] + \log\left(\sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}\right)\right\} \tag{12}$$

μ，τ^2の REML 推定量においても$l_{RL}(\mu,\tau^2)$を最大化する事により得られる。最大化は ML 推定量の算出と同様に初期値を与え，反復計算により REML 推定量が収束するまで繰り返す。REML 推定量はバイアス及び精度の面でバランスが良い推定量とされている（Viechtbauer 2005）。

2.1.5. その他

その他のτ^2の推定に関して，以下の推定量も有用であり，地域の数や地域間の異質性の程度から適宜使い分ける事も可能である。各推定法の詳細は付録に示す。

- Hedges-Olkin 推定量（Hedges and Olkin 1985）
- Paule-Mandel 推定量（Paule and Mandel 1989）
- Hunter-Schmidt 推定量（Schmidt and Hunter 2015）
- Sidik-Jonkman 推定量（Sidik and Jonkman 2005）
- 経験ベイズ推定量（Knapp and Hartung 2003）

2.2 従来の一貫性の評価方法

国際共同治験における地域間の治療効果の違いの評価に関しては，視覚的な評価としてフォレストプロットがあげられる。また，同時に治療効果と地域の交互作用の評価や厚生労働省から発出されている通知に基づく方法を用いて治療効果の地域間の一貫性の評価などが行われる。

2.2.1. フォレストプロット

フォレストプロットは治療効果の平均とその両側 95%信頼区間を地域別の結果と固定効果モデル及び変量効果モデルによる全体の結果を並べて表示する事により，各地域の治療効果の平均値が全体からどの程度乖離しているかを見ると同時に各地域の治療効果の信頼区間も見る事で偶然によるばらつきかどうかを評価する事ができる。例として，本研究における提案法を適用するために用意したリバーロキサバンの RECORD 試験の事例におけるフォレストプロットを図 2 に示す。事例の詳細については 3.5.1 節で示すため，ここでは記載しない。この事例でわかる通り，地域数が多い場合，各地域の患者数は少なくなり，両側 95%信頼区間は広くなる。治療効果の点推定値が全体集団の結果から外れていたとしても両側 95%信頼区間は重なり，どの地域が外れ値となっているかは判断が困難である。

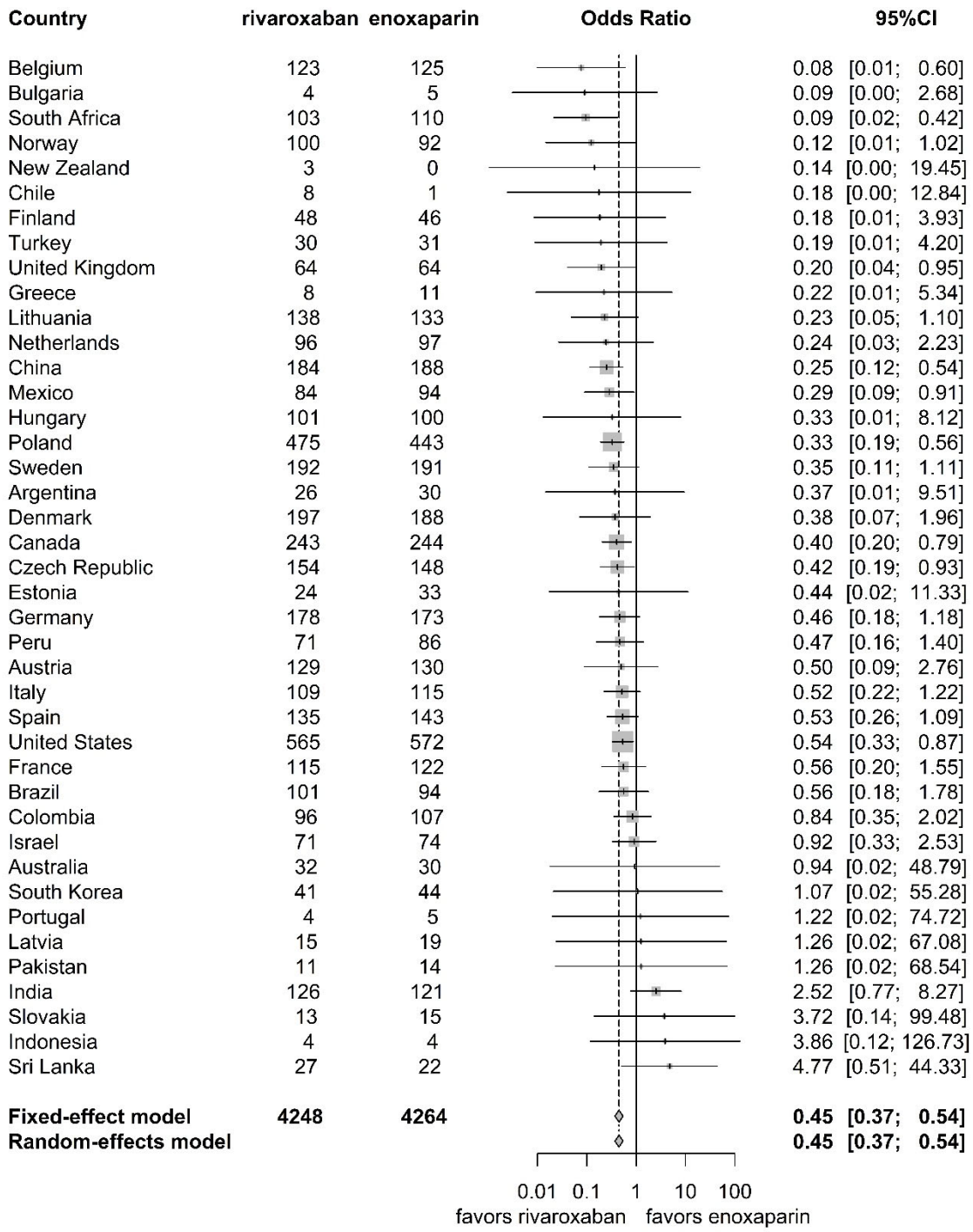

図 2 RECORD 試験における国別オッズ比に対するフォレストプロット

2.2.2. 治療効果と地域の交互作用の評価

治療効果と地域の交互作用を評価する方法の一つとして Cochran の Q 統計量がある（Chen et al. 2010）。Cochran の Q 統計量を考える時，帰無仮説を$H_0: \theta_1 = \theta_2 = \cdots =$

$\theta_k = \theta$とし，対立仮説に各地域の治療効果がいずれか一つでも異なるという仮説を立て，帰無仮説が棄却される場合に治療効果と地域に交互作用があると結論づける。これに対して，式(9)の Cochran の Q 統計量を考える。帰無仮説の下で Q 統計量は自由度$k-1$のχ^2分布に従うため，有意水準αのパーセント点を基準に検定を行う。

別の交互作用検定として質的交互作用を評価する Gail-Simon 検定がある（Gail and Simon 1985）。質的交互作用がないという事は，

$$O^+ = \{\theta_1 \geq 0, \cdots, \theta_k \geq 0\} \tag{13}$$

であるか，

$$O^- = \{\theta_1 \leq 0, \cdots, \theta_k \leq 0\} \tag{14}$$

である場合となる。Gail-Simon 検定は帰無仮説をH_0:すべてのθ_iに対して$\theta_i > 0$，又はすべてのθ_iに対して$\theta_i < 0$とし，対立仮説をH_0が成り立たない場合とし，これに対する尤度比検定となる。検定統計量は

$$Q_{GS} = \min(Q_{GS}^+, Q_{GS}^-) > c \tag{15}$$

で与えられる。この時，Q_{GS}^+とQ_{GS}^-は

$$Q_{GS}^+ = \sum_{i=1}^{k} \frac{y_i^2}{\sigma_i^2} I(y_i > 0)\,, Q_{GS}^- = \sum_{i=1}^{k} \frac{y_i^2}{\sigma_i^2} I(y_i < 0) \tag{16}$$

で与えられ，cは適当な棄却限界値，$I(\cdot)$は指示関数を表す。この検定統計量がχ^2分布の加重和に従うため，有意水準αのパーセント点を基準に検定を行う。

もし被験者の個別のデータが利用できる場合には回帰モデルにおいても交互作用を評価する事ができる。その場合，回帰モデルに治療効果と地域の交互作用項を加える。被験者数が多い場合には回帰モデルによる交互作用検定と Cochran の Q 統計量による検定の結果は同様の結果を示す。一方，個々の被験者の背景因子の違いを評価したい場合には，回帰モデルはそれらの因子を同時に調整する事ができる。

しかしながら，一般に国際共同治験では治療効果と地域の交互作用を検討する目的で試験がデザインされていないため，交互作用検定の検出力は低いと考えられる（Chen et al. 2010, Koshimizu 2003）。そこで，治療効果の地域間の異質性の大きさを評価するため，メタアナリシスで提案されている Cochran の Q 統計量を用いた Higgins のI^2統計量も有用である（Higgins et al. 2003）。Higgins のI^2統計量は以下の式で定義される。

$$I^2 = 100 \times \frac{Q - (k-1)}{Q} \tag{17}$$

負の値となる場合は 0 とし，0 から 100 の値をとる。0 に近いほど異質性がなく，大きい値ほど異質性がある。また，一般に以下の分類を目安に異質性の程度を判断できる。

- $0 \leq I^2 < 25$: 異質性なし
- $25 \leq I^2 < 50$: 低い異質性
- $50 \leq I^2 < 75$: 中程度の異質性
- $75 \leq I^2 \leq 100$: 高い異質性

ただし，これらの方法で交互作用や地域間の異質性を評価できたとしても，どの地域に交互作用があるのかを検出する事はできない。

2.2.3. 一貫性の評価方法

日本では厚生労働省から「国際共同治験における基本的考え方」の通知が発出されている（厚生労働省 2007）。この通知では全体集団と各地域の治療効果の一貫性評価のための方法及び各地域の被験者数設計に関する方法が 2 つ提案されている。

方法 1: プラセボ群と治験薬群での群間差をD，その場合の全集団での群間差をD_{all}，日本人集団における群間差をD_{Japan}とすると，$D_{Japan}/D_{all} > \pi$が成立するような確率が 80%以上となるように日本人被験者数を設定する。πについては，適切な値を設定する必要があるが，一般的には 0.5 以上の値をとる事が推奨される。この方法では，日本人被験者数を最小にしようとすると，全体での被験者数が増加し，全体での被験者数を最小にしようとすると日本人被験者数が増加するという関係が認められる。

方法 2: 全集団におけるプラセボ群と治験薬群での群間差をD_{all}，例えば 3 地域が試験に参加し，各地域でのプラセボ群と治験薬群での群間差をそれぞれD_1，D_2，D_3とすると，D_1，D_2，D_3が全て同様の傾向にある事を示す。例えばD_{all}が正の値をとるとすると，D_1，D_2，D_3のいずれの値も 0 を上回る確率が 80%以上となるように被験者数を設定する。この方法では，各地域から均等に被験者数を集積した場合に，確率が高くなるという傾向があり，全体の被験者数を変更する事なく日本人被験者数を検討する事が可能であるが，日本人の構成比率が小さく，被験者数が少ない場合に，地域間比較が十分に行えない場合がある事に留意すべきである。

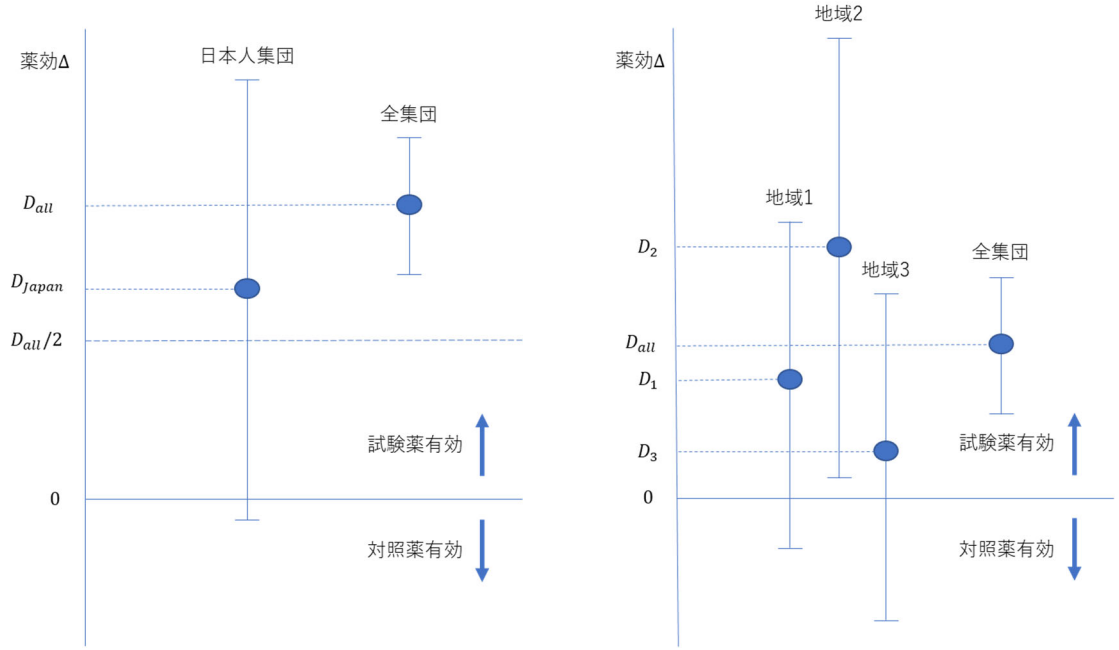

図 3 方法 1（左）と方法 2（右）の条件を満たす状況（日本製薬工業協会 2018）

方法 1 は以下を担保する必要がある。

$$\Pr\left(\frac{D_{Japan}}{D_{all}} > \pi\right) = 1 - \beta' \tag{18}$$

ここで，$1-\beta'$は正しく地域間の治療効果の一貫性を示せる確率を示し，$1-\beta' \geq 0.8$とする。ここで，Quan らは固定効果モデルの下で観測された治療効果の推定値を用いて式(18)を満たす閉形式を導いた（Quan et al. 2010a）。これにより，$\pi = 0.5$とし，評価項目が連続量の場合に各地域の被験者数を計算した時，少なくとも全体の 22.4%の被験者数が必要となる。よって，地域数が 5 地域以上の場合にはこの条件を満たす事ができない。また，方法 2 では河合らがすべての地域に対して真の治療効果が同じである時の固定効果モデルの下で一貫した結果を示す確率をシミュレーションによって考察した（Kawai et al. 2008）。そこでは，全体の検出力が 90%の時に 3 つの地域に対して方法 2 を用いると，地域間の治療効果の一貫性を示す確率を 80%以上担保するためには，各地域で全体の 15%以上を登録する必要がある事を示している。Quan らはこの方法 1 と方法 2 を統合した方法も提案している（Quan et al. 2010b）。これらの方法は明確に使い分けられておらず，疾患の希少性などを考慮して，規制当局と議論をした上で決定される。しかしながら，国際共同治験では患者の登録が各地域で競合的に行われるため，いずれの方法においても一貫性の評価に必要な被験者数が確保されない場合がある。一貫性の評価に必要な被験者数が確保されない場合，結果的に選択した方法の基準を満たしたとしても，偶然基準を満たしてしまった可能性が疑われてしまう。

3. 外れ値となる地域及び影響力のある地域の検出方法

本論文では，国際共同治験において外れ値となる地域及び影響力のある地域を効果的に検出できる一連の新しい影響診断ツールを提案する。最初に，固定効果モデル，及び変量効果モデルに対して，df ベータ型の影響尺度である leave-one-out cross-validation（LOOCV）により得られるスチューデント化残差を用いた方法を提案する。次に，固定効果モデル及び変量効果モデルに対して，移動平均モデルを用いた個々の地域に対する治療効果の不一致を評価するためのモデルベースの有意性検定を提案する。さらに，変量効果モデルに基づく影響度尺度として，（1）治療効果の全体分散に基づく相対的変化尺度及び（2）地域間分散に基づく相対的変化尺度を提案する。（1）と（2）のいずれも LOOCV で評価する。さらに，これらの影響尺度に対して基準を設定して統計的有意性を評価する際に，影響尺度のばらつきを考慮するため，影響尺度の分布の推定にパラメトリックブートストラップ法を適用する。

3.1 LOOCV 型のスチューデント化残差に基づく方法

最初に従来の回帰診断で用いられる df ベータ統計量に類似した LOOCV 型の統計量を提案する。ここで提案する方法は残差を用いた統計量となるが，残差に基づく指標はあらゆる解析尺度で比較可能とするため，残差の標準誤差で標準化する事によって定義する（Belsley et al. 1980）。これをスチューデント化残差と呼び，以下のように定義する。

$$r_i = \frac{y_i - \hat{\mu}}{\sqrt{Var[y_i - \hat{\mu}]}} \tag{19}$$

ここで，$Var[y_i - \hat{\mu}] = \hat{w}_i^{*-1} - \left(\sum_{i=1}^{k} \hat{w}_i^*\right)^{-1}$とし，$\hat{w}_i^* = \left(\hat{\tau}^2 + \sigma_i^2\right)^{-1}$とする。$\hat{\tau}^2$は地域間の分散の推定値，$\sigma_i^2$は各地域の分散の推定値とする。なお，ここでは$\hat{\tau}^2$，$\hat{\mu}$を変量効果モデルの式(2)に対する$\tau$，$\mu$の REML 推定値とするが，モデルパラメータの推定には 2.1 節で紹介した他の様々な方法を採用する事ができる。スチューデント化残差は各地域の治療効果y_iの全体平均の推定量からの逸脱を評価できるが，評価する地域自体の情報を用いて全体平均μを推定する。これにより，データの重複によって生じるバイアスである Optimism のある影響度を評価してしまう。そのため，評価したい地域と他の地域の乖離を評価する指標としては適切ではない。そこで，回帰分析の従来の影響診断では，LOOCV 型の尺度が Optimism を回避するために広く採用されている（Steyerberg et al. 2001）。ここでは，$\hat{\mu}^{(-i)}$及び$\hat{\tau}^{2(-i)}$をi番目（$i = 1,2,\ldots,k$）の地域を除外した$k-1$個の地域のデータに基づく変量効果モデルの式(2)における REML 推定値とする。この時，LOOCV 型のスチューデント化残差を以下のように定義する。

$$t_i = \frac{y_i - \hat{\mu}^{(-i)}}{\sqrt{Var[y_i - \hat{\mu}^{(-i)}]}} \tag{20}$$

ここで，$Var\left[y_i - \hat{\mu}^{(-i)}\right] = \left(\hat{w}_i^{*(-i)}\right)^{-1} + \left(\sum_{j \neq i} \hat{w}_j^{*(-i)}\right)^{-1}$とし，$\hat{w}_j^{*(-i)} = \left(\hat{\tau}^{2(-i)} + \sigma_j^2\right)^{-1}$，（$j = 1,2,\ldots,k$）とする。これにより，$t_i$は$i$番目の地域を除外した$k-1$個の地域のデータから推定された変量効果モデルの式(2)において，i番目の地域のデータの予測されたスチューデント化残差と解釈され，これは Optimism を含まない。このスチューデント化残差は異質性の分散の推定値$\hat{\tau}^{2(-i)}$と各地域の治療効果の分散の推定値σ_i^2の両方に依存する。したがって，y_iが$\hat{\mu}^{(-i)}$から離れていたとしても，$\hat{\tau}^{2(-i)}$やσ_i^2が十分大きければ，i番目の地域の治療効果は外れ値とは判断されない。

外れ値である地域と判断するため，判断基準となる閾値をt_iの標本分布によって得る事ができる。仮定した変量効果モデルの式(2)が正しい場合，t_iは標準正規分布に従う。基準となる閾値は任意に設定可能ではあるが，広く用いられている基準として両側 5%点を採用する場合は 1.96 をt_iの絶対値と比較する事ができる。この基準を満たす場合，その地域は偶然のばらつきの範囲を超える外れ値の可能性であると考えられる。しかし，この分布は大標本近似を仮定しているため，実際はばらつきが適切に定量化できていない可能性がある。そこで，実際のばらつきを考慮するため，t_iの標本分布の推定にパラメトリックブートストラップ法（Efron and Tibshirani 1994）を適用した。パラメトリックブートストラップ法のアルゴリズムは以下のように考えられる。

アルゴリズム 1（t_iの標本分布を推定するためのパラメトリックブートストラップ）

1. 観測されたデータに基づいて，帰無仮説の下で変量効果モデルの式(2)における REML 推定値$\hat{\mu}$及び$\hat{\tau}^2$を計算する。
2. $\hat{\mu}$と$\hat{\tau}^2$のパラメータを持つ式(2)の推定された分布$N\left(\hat{\mu}, \sigma_i^2 + \hat{\tau}^2\right)$から$B$回のパラメトリックブートストラップにより，$y_1^{(b)}, y_2^{(b)}, \ldots, y_k^{(b)}$（$b = 1,2,\cdots,B$）をリサンプリングする。

3. $\sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2$を持つ$b$番目のブートストラップサンプル$y_1^{(b)}, y_2^{(b)}, \ldots, y_k^{(b)}$に対して，LOOCV 型のスチューデント化残差$t_i^{(b)}$（$i = 1,2, \ldots, k$）を計算する。すべての$B$回のブートストラップサンプルについてこれを繰り返す。

4. $t_i^{(1)}, t_i^{(2)}, \ldots, t_i^{(B)}$の経験分布により$t_i$の標本分布のブートストラップ推定値を得る。

潜在的な外れ値となる地域を検出する基準を 5%とした場合，この経験分布における 2.5 パーセント点と 97.5 パーセント点は潜在的な外れ値となる地域を検出する閾値として利用できる。これにより検出された地域は偶然のばらつきの範囲を超える影響力のある外れ値であると考えられる。なお，上記の方法は，τ^2を 0 に固定し，μを共通の治療効果パラメータと解釈する事により，固定効果モデルにも同様に適用する事ができる。

3.2 尤度比検定に基づく方法

次に，Negeri と Beyene や野間らによりメタアナリシスの枠組みで提案された移動平均モデルを用いたモデルベースの尤度比検定を外れ値となる地域を検出する方法として提案する（Negeri and Beyene 2020, Noma et al. 2020）。移動平均モデルではある参加地域の治療効果が全体の治療効果と異なると仮定する。変量効果モデルの式(2)では，$k-1$個の地域の変量効果モデルの分布が$\theta_i \sim N(\mu, \tau^2)$であると仮定して，$j$番目（$i \neq j$）の地域の変量効果モデルの分布が$\theta_j \sim N(\mu + \zeta, \tau^2)$のように治療効果の全体平均が$\zeta$だけ移動していると仮定する。その上で以下の検定問題を検討する。

$$H_0: \zeta = 0 \text{ vs. } H_1: \zeta \neq 0 \tag{21}$$

この帰無仮説が棄却される時，j番目の地域の治療効果は全体の平均値から大きく乖離し，外れ値となる地域である可能性が疑われる。この移動平均は地域間の異質性τ^2によって特徴づける事ができない系統的な差を検出できる。

この検定問題に対して，尤度比検定を考えた時，帰無仮説の下での対数尤度関数は変量効果モデルの式(2)の対数尤度関数に対応しており，以下のように表される。

$$l_0(\mu, \tau^2) = -\frac{1}{2}\sum_{i=1}^{k}\left\{\log 2\pi(\sigma_i^2 + \tau^2) + \frac{(y_i - \mu)^2}{\sigma_i^2 + \tau^2}\right\} \tag{22}$$

さらに，対立仮説の下での対数尤度関数は以下の式で表される。

$$\begin{aligned} l_{1[j]}(\mu, \tau^2, \zeta) = &-\frac{1}{2}\left\{\log 2\pi(\sigma_j^2 + \tau^2) + \frac{(y_j - \mu - \zeta)^2}{\sigma_j^2 + \tau^2}\right\} \\ &-\frac{1}{2}\sum_{i \neq j}\left\{\log 2\pi(\sigma_i^2 + \tau^2) + \frac{(y_i - \mu)^2}{\sigma_i^2 + \tau^2}\right\} \end{aligned}$$

そして，尤度比統計量は以下のように与えられる。

$$T_{[j]} = -2\left\{l_0(\tilde{\mu}, \tilde{\tau}^2) - l_{1[j]}\left(\tilde{\mu}_{[j]}, \tilde{\tau}_{[j]}^2, \zeta_{[j]}\right)\right\} \tag{23}$$

ここで，$\tilde{\mu}$及び$\tilde{\tau}^2$は帰無仮説の下でのモデル式(2)の ML 推定値であり，$\tilde{\mu}_{[j]}$，$\tilde{\tau}^2_{[j]}$，$\zeta_{[j]}$はj番目の地域の移動平均モデルの ML 推定値である。尤度比統計量$T_{[j]}$は帰無仮説の下で自由度 1 のχ^2分布に従う。したがって，仮に 5%の有意水準の検定とした場合，外れ値と判断する基準として自由度 1 のχ^2分布の 95 パーセント点である 3.84 を採用する事ができる。なお，この提案法においても上記のモデルでτ^2を 0 に固定し，μを共通の治療効果パラメータとして解釈する事により，固定効果モデルに適用する事ができる。このモデルベースのアプローチにおける統計量は地域間の分散の ML 推定値$\tilde{\tau}^2_{[j]}$や各地域の治療効果の分散σ_j^2にも依存する。したがって，$\zeta_{[j]}$が大きかったとしても，$\tilde{\tau}^2_{[j]}$やσ_j^2が大きい場合には，j番目の地域の治療効果は外れ値とは判断されない。前節の議論と同様に，大標本近似は現実的な状況で起こりえない可能性があり，χ^2近似は有効でない場合がある（Noma et al. 2020, Veroniki et al. 2019）。よって，尤度比統計量$T_{[j]}$の標本分布の推定のため，パラメトリックブートストラップ法を適用する。

アルゴリズム 2（尤度比統計量の標本分布を推定するためのパラメトリックブートストラップ）

1. 観測されたデータに基づいて，ML 推定値$\tilde{\mu}$，$\tilde{\tau}^2$を算出する。
2. $\tilde{\mu}$と$\tilde{\tau}^2$のパラメータを持つ式(2)の推定された帰無仮説に基づくモデル$N\left(\tilde{\mu}, \sigma_i^2 + \tilde{\tau}^2\right)$から$B$回のパラメトリックブートストラップにより，$y_1^{(b)}, y_2^{(b)}, \ldots, y_k^{(b)}$（$b = 1,2,\cdots, B$）をリサンプリングする。
3. $\sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2$をもつ$b$番目のブートストラップサンプル$y_1^{(b)}, y_2^{(b)}, \ldots, y_k^{(b)}$から最尤推定値$\tilde{\mu}^{(b)}$，$\tilde{\tau}^{2(b)}$，及び$\tilde{\mu}_{[j]}^{(b)}$，$\tilde{\tau}_{[j]}^{2(b)}$，$\tilde{\zeta}_{[j]}^{(b)}$を計算し，以下の尤度比統計量
$$T_{[j]}^{(b)} = -2\left\{l_0\left(\tilde{\mu}^{(b)}, \tilde{\tau}^{2(b)}\right) - l_{1[j]}\left(\tilde{\mu}_{[j]}^{(b)}, \tilde{\tau}_{[j]}^{2(b)}, \zeta_{[j]}^{(b)}\right)\right\}$$
を算出する。これをB個のブートストラップサンプルに対して繰り返す。
4. $T_{[j]}^{(1)}, T_{[j]}^{(2)}, \ldots, T_{[j]}^{(B)}$の経験分布により$T_{[j]}$の標本分布のブートストラップ推定値を得る。

有意水準を 5%とした場合，この経験分布における 5 パーセント点を潜在的な外れ値となる地域を検出する閾値として利用でき，この経験分布上での実際のデータで算出した尤度比検定統計量のパーセント点を用いてブートストラップ P 値を計算できる。

3.3 全体分散の相対的変化に基づく方法

外れ値となる地域の検出及び影響診断のアプローチとして，LOOCV 型の全体分散の推定値の相対的変化を評価する方法を提案する。Viechtbauer と Cheung はメタアナリシスの枠組みでこの方法を提案した（Viechtbauer and Cheung 2010）。これは変量効果モデル(2)に基づいて国際共同治験の枠組みに直接適用する事ができる。j番目の地域に対する外れ値の検出及び影響診断の統計量として，以下が与えられる。

$$VRATIO_j = \frac{Var\left[\hat{\mu}^{(-j)}\right]}{Var[\hat{\mu}]} = \frac{\sum_{i=1}^{k} \hat{w}_i^*}{\sum_{i \neq j} \hat{w}_i^{*(-j)}} \tag{24}$$

この統計量では一つの地域を除いたデータと全ての参加地域を含むデータの治療効果の全体分散の相対的変化を評価する。$VRATIO_j$は治療効果の推定値の分散に対してj番目の地域の影響を示し，0 から∞の値をとる。1 の周辺の値をとる場合，j番目の地域が除外されても，分散の推定値の変化は少なく，除外した地域が全体にあまり影響していない事を表す。$VRATIO_j$が 1 より大きい場合，j番目の地域の除外は全体の推定値のばらつきを大きくしている。しかしながら，j番目の地域を除外する事により被験者数が減少し，ばらつきが大きくなるのは当然と考えられる。従って，そのような地域は外れ値として解釈されない。逆に，$VRATIO_j$が 1 より小さい時，j番目の地域を含む事によって被験者数が増加し精度が増しているにも関わらず，全体の推定値のばらつきを大きくしている。これはj番目の地域が地域間の異質性τ^2を増大させているという事を意味する。この場合，j番目の地域が外れ値となり，全体集団から外れている可能性がある。したがって，最も小さい$VRATIO_j$を持つ地域は外れ値であり，影響力のある地域であると考えられる。明確な基準により外れ値を特定するため，$VRATIO_j$の分布を推定し，基準を設定する必要がある。そこで，3.1 節のパラメトリックブートストラップ法のアルゴリズム 1 のt_iを$VRATIO_j$に置き換える事で$VRATIO_j$の標本分布を推定する事ができる。潜在的な外れ値となる地域を検出する基準を 5%とすると，パラメトリックブートストラップ法で生成した分布の下位 5 パーセント点を基準として利用できる。

固定効果モデルを用いた場合，治療効果の全体分散を用いた統計量はσ_i^2のみで構成される統計量となり，単純に個々の地域の被験者数を反映してしまう。従って，それは外れ値となる地域を検出するための適切な尺度ではないため，変量効果モデルを用いた解析のみに適用すべきである。

3.4 地域間分散の相対的変化に基づく方法

最後に，$VRATIO_j$と同様の影響尺度を変量効果モデルの式(2)の地域間分散の推定値に対して提案する。Viechtbauer と Cheung は，以下の一つの地域を除外したデータ及び全地域のデータに対する地域間分散の推定値の相対的変化を評価した方法を提案した（Viechtbauer and Cheung 2010）。

$$TRATIO_j = \frac{\hat{\tau}^{2(-j)}}{\hat{\tau}^2} \tag{25}$$

$TRATIO_j$の値も 0 から∞の値をとり，$VRATIO_j$と同様に解釈できる。$TRATIO_j$が 1 に近い場合，j番目の地域が除外されても，地域間分散の推定値の変化は少なく，除外した地域が全体にあまり影響していない事を表す。また，$TRATIO_j$が 1 より大きい場合，通常，地域数が減少する事により地域間分散が上昇する事は当然と考えられるため，外れ値である地域とは解釈しない。一方，$TRATIO_j$が 1 より小さい場合，j番目の地域の除外は地域間の異質性を減少させるため，外れ値となる治療効果を持つ地域と解釈する事ができる。外れ値となる治療効果を持つ地域の判定基準を決定するために，パ

ラメトリックブートストラップ法を用いて，3.1 節のアルゴリズム 1 でt_iを$TRATIO_i$に置き換える事により$TRATIO_i$の標本分布を推定する。仮に潜在的な外れ値となる地域を検出する基準を 5%とすると，ブートストラップ分布の下位 5 パーセント点を基準として利用できる。なお，この指標は固定効果モデルでは$\tau^2 = 0$となり，計算できないため，変量効果モデルの解析にのみ適用できる。Viechtbauer と Cheung によって議論されているように，相当な異質性が存在する場合に，この尺度は効果的な診断ツールとなる（Viechtbauer and Cheung 2010）。

3.5 事例解析

ここではリバーロキサバン，ロサルタン，メトプロロールの 3 つ薬剤に対してそれぞれ行われた国際共同治験の事例への応用を通じて提案法の有用性を検証した。

3.5.1. RECORD 試験

最初に人工股膝関節置換術を実施した患者を対象に深部静脈血栓症及び肺血栓塞栓における静脈血栓塞栓症の予防に対するリバーロキサバンの国際共同治験である RECORD 試験（FDA 2009, Turpie et al. 2011）の事例を用いる。この試験は 4 つの試験が実施され，RECORD-1 試験と RECORD-2 試験は人工股関節置換術を実施した患者を対象としており，RECORD-3 試験と RECORD-4 試験は人工膝関節置換術を実施した患者を対象としている。いずれも静脈血栓塞栓症の予防に対して，リバーロキサバンとエノキサパリンを比較しており，主要評価項目は静脈血栓塞栓症（症候性及び無症候性深部静脈血栓症，非致死性肺塞栓症，全死亡）の発現割合としている。いずれの試験においてもリバーロキサバンのエノキサパリンに対する優越性が示されている。アメリカの承認審査においては，この 4 つの国際共同治験のデータを併合し，各試験でエノキサパリンの投与期間や用量などが異なるため，試験間で共通してエノキサパリンが投与されている初回投与後 10 日から 14 日での静脈血栓塞栓症に対して，有効性の評価を行っている。本論文でも併合データに対して提案法を適用する。RECORD 試験の統合解析では，41 ヵ国が参加しており，リバーロキサバン群 4248 名，エノキサパリン群 4264 名が解析対象となった。主要評価項目の静脈血栓塞栓症はリバーロキサバン群で 181 名（4.3%），エノキサパリン群で 402 名（9.4%）に発現した。各国の治療効果のオッズ比とその両側 95%信頼区間を表すフォレストプロットは 2.2.1 節に示した図 2 の通りである。地域間分散τ^2を DerSimonian-Laird 推定量で推定した場合，$\hat{\tau}^2 = 0$となるため，固定効果モデルと変量効果モデルによる全体の治療効果とその両側 95%信頼区間は同じであり，0.45 [0.37; 0.54]となる。各地域の治療効果のオッズ比の点推定値は 0.08 から 4.77 の範囲にあり，両側 95%信頼区間も考慮すると，全体の治療効果から極端に外れている地域，または地域のクラスターなどがあるようには見えなかった。このように参加国数が多い場合，全体の治療効果からの差や参加国間の治療効果のばらつきは視覚的に確認できるが，各国の治療効果のばらつきが異なり，両側 95%信頼区間が重なるため，外れ値となる地域や影響力のある地域の特定は困難である。

治療効果と国の交互作用を評価するための Cochran の Q 統計量では有意な交互作用は確認されず（p=0.533），Gail-Simon 検定においても有意な質的交互作用は確認されなかった（p=0.997）。また，Higgins のI^2統計量は 0%であった。これらの結果より治

療効果と地域の交互作用はないと判断されるが，国の数も多く，各地域で被験者数が十分ではないため，交互作用を特定するための検出力が不足していると考えられる。

さらに，厚生労働省から発出されている通知「国際共同治験における基本的考え方」（厚生労働省 2007）に記載されている方法 1 と方法 2 に基づく評価について考える。

まず，方法 1 において全体の治療効果$D_{all} = -\log(0.45)$とし，各国の治療効果$D_i = -\log(\hat{\theta}_i)$に対して，$D_i/D_{all} > \pi$を基準に治療効果が一貫しているかどうかを考える。ここでは，$\pi = 0.5$とすると，コロンビア，イスラエル，オーストラリア，韓国，ポルトガル，ラトビア，パキスタン，インド，スロバキア，インドネシア，スリランカの 11 ヵ国がこの基準を満たしていないため，全体集団の結果と一貫していない治療効果を持つ国となる。ここで，方法 1 で定められている通り，式(18)に対して，上記の評価と同様に$\pi = 0.5$とし，正しく地域間の治療効果の一貫性を示せる確率を 80%とした場合の必要被験者数は全体の 22.4%となるので，1907 名が必要となる。そのため，この方法に基づくと先ほど挙げた全体集団と一貫していないいずれの国でも被験者数不足により結論づけられない事となる。

次に，方法 2 においても全体の治療効果$D_{all} = -\log(0.45)$とし，各国の治療効果$D_i = -\log(\hat{\theta}_i)$に対して，全ての国の治療効果$D_1, D_2, \cdots, D_{41}$が 0 を上回るという基準を用いて一貫した結果であるかどうかを考える。この時，韓国，ポルトガル，ラトビア，パキスタン，インド，スロバキア，インドネシア，スリランカが 0 を下回ってしまうため，国間で一貫した治療効果を持つとは言えない。方法 2 に関しても上記の基準に対して，正しく地域間の治療効果の一貫性を示せる確率を 80%以上とし，必要被験者数を計算すると，各地域で全体の 15%以上を登録する必要があり，1277 名が必要となる。この場合でも治療効果D_iが 0 を下回るいずれの国でも必要被験者数を満たしている国はなく，被験者数不足により結論づけられない事となる。また，方法 2 における必要被験者数は方法 1 の必要被験者数よりも少なくなってはいるが，依然として各国でこの必要被験者数を登録するのは困難であると考えられる。

そこで，本研究における提案法を RECORD 試験のデータに適用した。変量効果モデルに対して DerSimonian-Laird 推定量によって推定した地域間の異質性$\hat{\tau}^2$が 0 であるため，この事例では固定効果モデルのみを用いて，LOOCV 型のスチューデント化残差及び尤度比検定に基づく方法のみで評価した。なお，ここでは共通して外れ値であると判断する基準は 5%とした。また，パラメトリックブートストラップ法におけるリサンプリングは 2400 回とした。表 1 に LOOCV 型のスチューデント化残差t_iの絶対値が大きい地域の上位 10 地域のみを示し，パラメトリックブートストラップ法により推定したt_iの標本分布の 2.5%点と 97.5%点を外れ値と判断する基準として提示した。全体的に閾値は$N(0,1)$の 2.5 及び 97.5%点に対応する± 1.96と大きく異ならなかったが，いくつかはわずかに異なる値となった。この結果，インドのスチューデント化残差が 2.886 となり，パラメトリックブートストラップ法による標本分布の 97.5%点の 1.921 を大きく超えた値となった。続いて，スリランカのスチューデント化残差は 2.087 となり，パラメトリックブートストラップ法による標本分布の 97.5%点の 1.983 からわずかに超えた値となった。そのため，インドとスリランカは治療効果が無効の方向に外れている可能性がある。さらに，南アフリカのスチューデント化残差は−2.067 となり，パラメトリックブートストラップ法による標本分布の 2.5%点の−1.963 をわずかに下

回った値となった。そのため，南アフリカは治療効果が有効な方向に外れている可能性がある。

表 1 RECORD 試験における固定効果モデルの LOOCV 型のスチューデント化残差を用いた外れ値となる国の評価結果: t_iの絶対値の大きさに対する上位 10 ヵ国

国	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
インド	**2.886**	**−1.912**	**1.921**
スリランカ	**2.087**	**−2.024**	**1.983**
南アフリカ	**−2.067**	**−1.963**	**1.976**
ベルギー	−1.684	−1.920	1.925
中国	−1.524	−1.910	1.899
コロンビア	1.443	−2.005	1.854
イスラエル	1.409	−1.983	1.976
スロバキア	1.264	−1.882	2.009
インドネシア	1.210	−1.957	2.048
ノルウェー	−1.206	−1.916	1.985

各国の固定効果モデルにおける尤度比統計量を表 2 に示した。さらに，外れ値と判断する基準をパラメトリックブートストラップ法による尤度比統計量の標本分布の 95%点とし，対応する P 値も示した。また，ここではパラメトリックブートストラップ法により算出した P 値が小さい 10 ヵ国のみを提示した。この解析においてもインドの尤度比統計量は 8.332 となり，パラメトリックブートストラップ法による標本分布の 95%点である 3.719 から大きく上回っていた（p=0.003）。また，スリランカの尤度比統計量は 4.357 となり，パラメトリックブートストラップ法による基準値の 3.888 をわずかに上回っていた（p=0.039）。南アフリカにおいても尤度比統計量は 4.274 となり，基準値である 4.075 をわずかに上回っていた（p=0.043）。

表 2 RECORD 試験における固定効果モデルの尤度比統計量を用いた外れ値となる国の評価結果: パラメトリックブートストラップ法による P 値が小さい 10 カ国

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
インド	**8.332**	**3.719**	**0.003**
スリランカ	**4.357**	**3.888**	**0.039**
南アフリカ	**4.274**	**4.075**	**0.043**
ベルギー	2.836	3.942	0.095
中国	2.322	3.696	0.114
コロンビア	2.081	3.839	0.155
イスラエル	1.984	3.971	0.160
スロバキア	1.598	3.873	0.195
ノルウェー	1.453	3.809	0.229
ポーランド	1.443	3.794	0.231

以上の結果より，インド，スリランカ，南アフリカが外れ値となる治療効果を持つ国である可能性があると考えられる。また，異なる二つの方法において，上位 10 ヵ国に含まれている国は順位も含め概ね同じ結果が得られたため，より頑健な結果であると考えられる。アメリカの審査において，インドは 100 名超の被験者規模でありながら，オッズ比が 1.0 を上回っているが，静脈血栓塞栓症のイベント数も少なく信頼区間が広かったという事が言及されており，それ以上の議論については行われていない。しかしながら，インドとスリランカにおいては地理的にも近く，図 1 に示すような外因性や内因性が治療効果に影響していなかったか更なる調査が必要であったかもしれない。仮に何かの要因が治療効果に影響していると仮定し，主要な解析結果の感度解析としてインド，スリランカ，南アフリカを除外して解析した。その結果，オッズ比とその両側 95%信頼区間は 0.43 [0.36, 0.52]であり，治療効果の推定値がわずかに増加する結果となったが，全体的な結論は変わらないため，主要な解析結果の頑健性が示された。この結果より提案法は各地域の被験者数が限定されている場合でも，外れ値を客観的に評価し，影響を定量化する事ができ，外れ値となる治療効果を持つ地域を検出するために効果的に適用する事ができると考えられた。

3.5.2. RENAAL 試験

二つ目の事例として，2 型糖尿病性腎症患者におけるロサルタンの国際共同治験である RENAAL 試験の事例（Brenner et al. 2001, 独立行政法人医薬品医療機器総合機構 2006）において提案法を適用した。RENAAL 試験では 28 ヵ国が参加し，最終解析にはロサルタン群の 751 名とプラセボ群の 762 名が含まれた。主要評価項目は血清クレアチニン濃度倍増，末期腎不全，死亡から成る複合評価項目であり，ロサルタン群の 32

7 名（43.5%）とプラセボ群の 359 名（47.1%）に主要評価項目に該当するイベントが発現した。各国の主要評価項目のハザード比とその両側 95%信頼区間は図 4 のフォレストプロットに示した。なお，フランス，オランダ，ニュージーランド，スロバキア，カナダの 5 ヵ国はロサルタン群又はプラセボ群でイベントを発現していない事よりハザード比が推定不能となってしまうため，解析からは除外した。このフォレストプロットより各国の治療効果は 0.25 から 2.29 の範囲にあり，アメリカの患者数が大半を占めている事がわかるが，各国の両側 95%信頼区間は概ね重なっており，極端に外れている地域や地域のクラスターの存在を確認する事は困難である。地域間分散τ^2の DerSimonian-Laird 推定量は 0 であり，ハザード比とその両側 95%信頼区間は固定効果モデルと変量効果モデルで同一となり，0.87 [0.75, 1.02]であった。

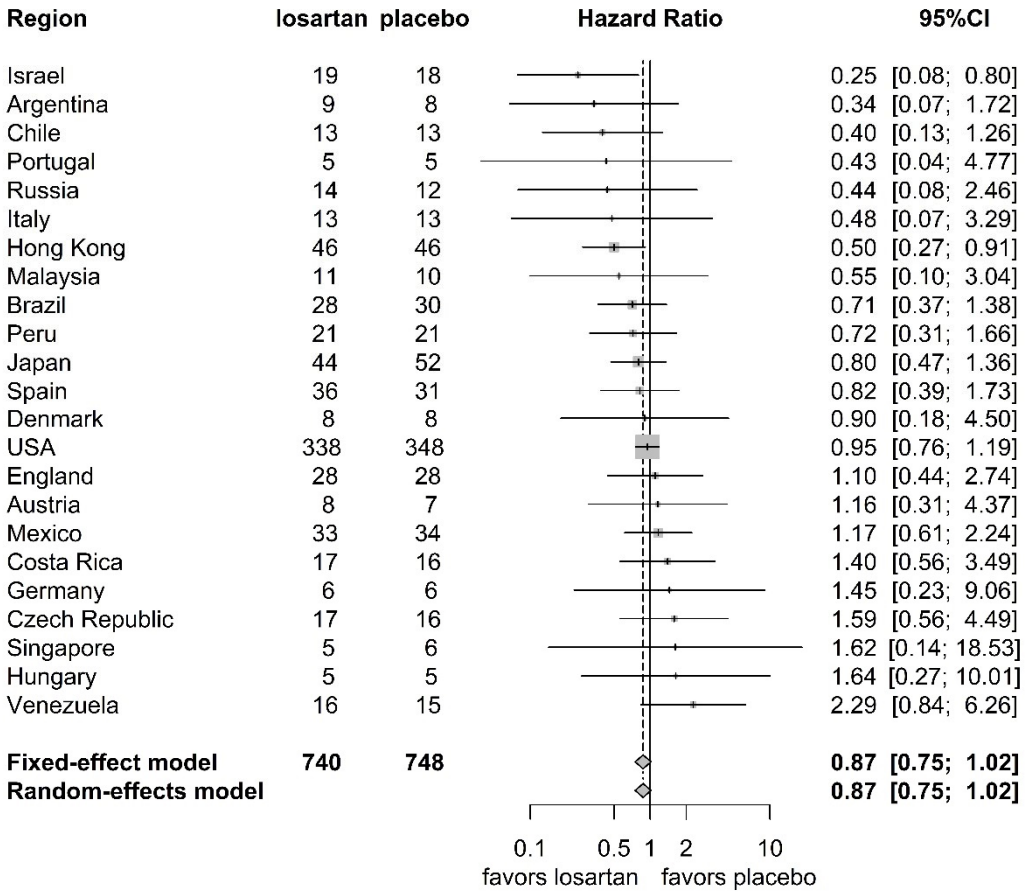

図 4 RENAAL 試験のおける国別ハザード比のフォレストプロット

治療効果と国の交互作用を評価するための Cochran の Q 統計量では有意な交互作用は確認されず（p=0.48），Gail-Simon 検定においても有意な質的交互作用は確認されなかった（p=0.906）。また，Higgins のI^2統計量は 0%であった。これらの結果より治療効果と国の交互作用はないと判断されるが，リバーロキサバンの事例と同様に国の数も多く，各地域で十分な被験者数がないため，交互作用を特定するための検出力が不足していると考えられる。

次に，厚生労働省の通知（厚生労働省 2007）における方法 1 と方法 2 に基づく評価について考える。全体の治療効果$D_{all} = -\log(0.87)$とし，各国の治療効果$D_i = -\log(\hat{\theta}_i)$とすると，方法 1 では基準$D_i/D_{all} > \pi$に対して，$\pi = 0.5$の場合にアメリカ，イギリス，オーストリア，メキシコ，コスタリカ，ドイツ，チェコ共和国，シンガポール，ハンガリー，ベネズエラの 10 ヵ国は基準を満たさなかった。そのため，これらの国は治療効果が全体集団と一貫していないという事になる。また，リバーロキサバンの事例でも方法 1 に基づいた各国の必要被験者数を計算すると，全体の 22.4%である 339 名が各国で必要となる。そのため，この方法に基づくと先ほど挙げた全体集団と一貫していない国のうち，アメリカのみが必要被験者数を満たしている事となり，アメリカだけは一貫していない地域として結論付ける事ができるが，その他の国では被験者数不足により結論付ける事ができない。また，28 ヵ国全てで 339 名を登録する場合，全体の必要被験者数を超える事となり，現実的ではない。

同様に方法 2 の場合では，$D_i > 0$の基準を満たさない国はイギリス，オーストリア，メキシコ，コスタリカ，ドイツ，チェコ共和国，シンガポール，ハンガリー，ベネズエラとなる。方法 2 に基づく必要被験者数を計算すると，全体の 15%である 227 名が各国で必要となるが，一貫性の基準を満たさない国のいずれもこの被験者数を満たしていない。この必要被験者数も 28 ヵ国全てで 227 名を登録する事は現実的ではない。

これに対して提案法を適用する。このデータにおいても最初に述べた通り，DerSimonian-Laird 推定量によって推定した地域間の異質性$\hat{\tau}^2$が 0 であるため，固定効果モデルのみを用いて，LOOCV 型のスチューデント化残差と尤度比検定に基づく方法のみで評価した。なお，ここでは共通して外れ値であると判断する基準は 5%とした。また，パラメトリックブートストラップ法におけるリサンプリングは 2400 回とした。

まず，LOOCV 型のスチューデント化残差に基づく方法による結果を表 3 に示した。パラメトリックブートストラップ法による 2.5%点及び 97.5%点を基準とした場合，イスラエルの LOOCV 型のスチューデント化残差は－2.126 であり，パラメトリックブートストラップ法による標本分布の 2.5%点である－1.984 よりやや下回った値となった。そのため，イスラエルは治療効果が有効な方向に外れている可能性がある。

表 3 RENAAL 試験における固定効果モデルの LOOCV 型のスチューデント化残差を用いた外れ値となる国の評価結果: t_iの絶対値の大きさに対する上位 10 ヵ国

国	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
イスラエル	**−2.126**	**−1.984**	**1.921**
ベネズエラ	1.901	−1.983	2.085
香港	−1.879	−2.047	2.017
チリ	−1.342	−2.014	1.947
アルゼンチン	−1.147	−1.994	1.913
コスタリカ	1.1437	−1.916	1.918
チェコ共和国	1.026	−1.884	1.926
アメリカ	0.970	−1.951	2.046
メキシコ	0.906	−1.968	1.937
ロシア	−0.784	−1.969	1.997

次に，各国の固定効果モデルにおける尤度比統計量を表 4 に示した。また，外れ値と判断する基準をパラメトリックブートストラップ法による標本分布の 95%点として，対応する P 値も示した。ここでも P 値が小さい上位 10 ヵ国のみを提示した。その結果，イスラエルは尤度比統計量 4.521 となり，パラメトリックブートストラップ法による標本分布の 95%点である 3.650 を上回っていた（p=0.033）。さらに，香港の尤度比統計量は 3.533 となり，パラメトリックブートストラップ法による標本分布の 95%点である 3.533 と等しかった（p=0.500）。

表 4 RENAAL 試験における固定効果モデルの尤度比統計量を用いた外れ値となる国の評価結果: パラメトリックブートストラップ法による P 値が小さい 10 ヵ国

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
イスラエル	**4.521**	**3.650**	**0.033**
香港	**3.533**	**3.533**	**0.050**
ベネズエラ	3.615	3.905	0.059
チリ	1.801	3.690	0.172
チェコ共和国	1.308	3.909	0.261
アルゼンチン	1.315	3.838	0.264
コスタリカ	1.053	3.970	0.297
アメリカ	0.942	3.521	0.336
メキシコ	0.820	3.828	0.373
ロシア	0.615	3.986	0.426

以上の結果より，イスラエルと香港が外れ値となる治療効果を持つ国である可能性が考えられる。さらに，日本の承認審査のおいては各地域の被験者規模を鑑みて，地理的に近い 4 つの併合地域（アジア地域，欧州地域，中南米地域，北米地域）の部分集団でも検討されている。アジア地域はイスラエルと香港を含み，同じ要因により外れ値となる治療効果を持つ国であったかもしれないと考えられる。そこで，併合地域に関しても外れ値の評価を行った。日米 EU 医薬品規制調和国際会議ガイドライン（International Conference ICH 2017）には地域の併合は規制当局の意思決定を助ける事につながるかもしれないと述べられており，併合された地域の影響を定量化する事は有用であると考えられる。まず，併合地域別のハザード比の推定値を表すフォレストプロットを図 5 に示した。この結果より，中南米地域，欧州地域，北米地域は 0.91 から 0.95 の範囲であり，アジア地域は 0.54 と大きな治療効果があり，外れ値となる地域である可能性が高い。しかしながら，各地域の被験者規模も異なるため，これだけでアジア地域が外れ値となる地域であると判断する事は困難である。なお，地域間分散τ^2の DerSimonian-Laird 推定量は 0.04 であり，固定効果モデルによるハザード比の推定値とその両側 95%信頼区間は 0.85 [0.73; 0.99]となり，変量効果モデルによるハザード比の推定値とその両側 95%信頼区間は 0.82 [0.64; 1.06]となった。

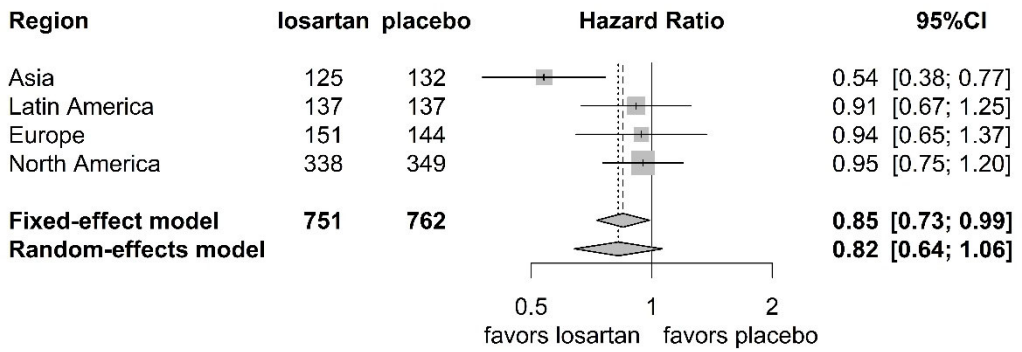

図 5 RENAAL 試験における地域別ハザード比のフォレストプロット

治療効果と地域の交互作用を評価するために Cochran の Q 統計量を考えると，有意な交互作用が確認されたが（p=0.049），Gail-Simon 検定では有意な質的交互作用は確認されなかった（p=0.875）。また，Higgins のI^2統計量は 61.8%となり，中等度の異質性が示唆された。

厚生労働省の通知に基づいた評価について考えると，この通知では本来治療効果が無効な方向に対して，全体集団と各地域の治療効果が一貫していない事を示すものとなっており，この事例のようにアジア地域のみが有効な方向へ外れている場合には適していないと考えられる。

そこで，提案法を用いる事によりアジア地域が外れ値であり，影響力のある地域である事を確かめる。前述したとおり，変量効果モデルにおいて DerSimonian-Laird 推定量による地域間分散の推定値$\hat{\tau}^2$が 0.04 となり，小さな異質性が確認されたため，4 つの提案法を用いて検討を行った。なお，ここでは共通して外れ値であると判断する基準は 5%とした。また，パラメトリックブートストラップ法におけるリサンプリングは 2400 回とした。最初に，固定効果モデルにおける LOOCV 型のスチューデント化残差を用いた方法による結果を表 5 に示し，変量効果モデルにおける LOOCV 型のスチューデント化残差を表 6 に示す。その結果，固定効果モデルにおいて，全体的に閾値が$N(0,1)$の 2.5 及び 97.5%点に対応する± 1.96と大きく異ならなかったが，変量効果モデルにおいては大きく異なる値となり，地域間の異質性が影響したものと考えられる。この基準に基づき，アジアのスチューデント化残差t_iは固定効果モデルで−2.796，変量効果モデルで−2.792 となり，パラメトリックブートストラップ法によるt_iの標本分布における 2.5%点（固定効果モデルで−2.058，変量効果モデルで−2.478）をいずれも下回った。

表 5 RENAAL 試験の固定効果モデルにおける LOOCV 型のスチューデント化残差を用いた外れ値となる地域の評価結果

地域	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
アジア	**−2.796**	**−2.058**	**1.991**
北米	1.290	−1.924	1.912
欧州	0.599	−2.032	2.002
中南米	0.532	−1.907	1.921

表 6 RENAAL 試験の変量効果モデルにおける LOOCV 型のスチューデント化残差を用いた外れ値となる地域の評価結果

地域	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
アジア	**−2.792**	**−2.478**	**2.561**
北米	0.612	−2.808	2.873
欧州	0.475	−2.421	2.334
中南米	0.389	−2.810	2.583

次に，表 7，表 8 に尤度比検定に基づく方法による結果を示す。この解析においてもアジアの尤度比統計量が固定効果モデルで 7.815，変量効果モデルで 6.935 となり，パラメトリックブートストラップ法による尤度比統計量の標本分布における 95%点（固定効果モデルで 3.889，変量効果モデルで 5.013）を大きく上回った。

表 7 RENAAL 試験の固定効果モデルに対する尤度比統計量を用いた外れ値となる地域の評価結果

地域	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
アジア	**7.815**	**3.889**	**0.009**
北米	1.663	3.725	0.182
欧州	0.358	4.090	0.544
中南米	0.283	3.676	0.608

表 8 RENAAL 試験の変量効果モデルに対する尤度比統計量を用いた方法を用いた外れ値となる地域の評価結果

地域	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
アジア	**6.935**	**5.013**	**0.014**
北米	0.899	5.381	0.476
欧州	0.339	5.035	0.624
中南米	0.257	4.836	0.651

次に，全体分散の推定値及び地域間分散の推定値の相対的変化を評価した解析結果をそれぞれ表 9，表 10 に示した。また，影響力のある地域と判断する基準としてパラメトリックブートストラップ法における$VRATIO$の標本分布の 5%点も併せて示した。全体分散の相対的変化に基づく方法ではアジアで$VRATIO$が 0.418 となり，基準値 0.461 を下回るため，影響力のある地域と考えられる。また，地域間分散の相対的変化に基づく方法でもアジアで$TRATIO$が 0 となり，基準値 0.003 を下回るため，同様に影響力のある地域と考えられる。$VRATIO$と$TRATIO$を影響力の指標と考えても，アジアのみが 1 を大きく下回っており，影響力の高い地域である事がわかる。

表 9 RENAAL 試験における全体分散の相対的変化を評価した外れ値となる地域の評価結果

地域	$VRATIO$	ブートストラップ 5%点
アジア	**0.418**	**0.461**
欧州	1.788	0.448
北米	1.877	0.521
中南米	1.954	0.462

表 10 RENAAL 試験における地域間分散の相対的変化を評価した外れ値となる地域の評価結果

地域	$TRATIO$	ブートストラップ 5%点
アジア	**0.000**	**0.003**
北米	1.529	0.001
欧州	1.608	0.002
中南米	1.733	0.002

以上より，全体としてアジア地域だけが提案した 4 つの方法において，一貫して偶然のばらつきの範囲を超える外れ値をもつ地域として検出された。独立行政法人医薬品医療機器総合機構の審査報告書（独立行政法人医薬品医療機器総合機構 2006）によると，地域間の異質性については日本の承認審査時にも議論された。この報告書ではイスラエルではなく，香港の治療効果が全体の治療効果に影響を与えたという申請者の見解が記載されている。また，地域間の治療効果の差が生じた要因として，ベースライン尿中アルブミン/クレアチニン比の影響が考えられた。さらに，治療効果の投与中止率も地域間の治療効果の差が生じた要因と考えられた。なぜなら，治験薬の投与が中止された場合，患者が来院しなくなる事により血清クレアチニン値倍増のデータが収集できなくなり，複合評価項目による評価はロサルタンの治療効果を必ずしも適切に反映しない可能性があるためである。特にアメリカは治験薬の投与中止率が最も高く，治験薬の投与期間も短かった。これは，北米において試験中に糖尿病性腎症及び高血圧の治療に関するガイドラインが公表され，ガイドラインで推奨された ACE 阻害剤に切り替えられた事により治験薬の中止率が高かった可能性がある。しかしながら，結論としては事後的な部分集団解析を行うには被験者数が不足している事から，この試験から得られる情報には一定の限界があると考えられた。

以上の考察から仮に何かの要因によりアジア地域の治療効果が外れ値であると仮定し，感度分析としてアジア地域以外の治療効果を確認するため，全体からアジア地域を除いた解析を行った結果，固定効果モデルのハザード比の推定値とその両側 95%信頼区間は 0.94 [0.79; 1.11]となった。また，変量効果モデルにおける地域間の異質性の推定値は0になったため，変量効果モデルは固定効果モデルと同一の結果を示した。よって，アジア地域を除外する事によりハザード比の推定値は大きく変化し，有意差は消失した。これにより，アジア地域はこの国際共同治験の全体的な治療効果の推定と結論に強い影響を及ぼしていたと推測できる。アジア地域とそれ以外の地域の治療効果の推定値の違いは直観的に明らかであるが，提案法を用いる事によってある統計的基準に基づいてアジア地域が外れ値である地域として検出し，影響力のある地域として特定できたと考えられる。過去の報告では地域別の部分集団解析のみが提示されており，直観的にアジア地域が異なる治療効果を有する地域と考えられたが，異なる治療効果であると判断するための情報は限られていたという結論に至っていた。一方，提案法は外れ値の検出において，客観的評価を提供でき，外れ値となる地域の影響を評価できた。

3.5.3. MERIT-HF 試験

最後の事例解析として，メトプロロール徐放錠の慢性心不全患者に対する国際共同治験である MERIT-HF 試験（MERIT-HF Study Group 1999, Hjalmarson et al. 2000）においても 4 つの提案法を適用した。MERIT-HF 試験では 14 ヵ国が参加し，メトプロロール群 1990 名，プラセボ群 2001 名のデータが解析された。主要評価項目は二つ設定されており，一つは全死亡であり，もう一つは全死亡及び理由を問わない入院が設定されていた。ここでは全死亡で一貫していない地域の存在が疑われていたため，全死亡に焦点を当てて考察する事とし，全死亡の有無に対するメトプロロール群とプラセボ

群のオッズ比で評価した。各国のオッズ比の推定値を表すフォレストプロットを図 6 に示した。ここで，フィンランドとスイスにおいてはいずれもメトプロロール群又はプラセボ群で死亡した被験者が存在していない事よりオッズ比が推定不能となってしまうため，解析からは除外した。このフォレストプロットより，各国のオッズ比の点推定値は 0.19 から 1.18 の範囲にあり，アメリカが全体の約 25%を占める事がわかる。また，固定効果モデルを用いた全体の解析ではオッズ比の推定値とその両側 95%信頼区間は 0.67 [0.54; 0.84]であり，有意な結果が示された。変量効果モデルでは地域間の異質性に対する DerSimonian-Laird 推定量が$\hat{\tau}^2 = 0.0616$となり，小さな異質性が確認され，オッズ比の推定値とその両側 95%信頼区間は 0.63 [0.47; 0.83]であり，同様に有意な結果が示された。

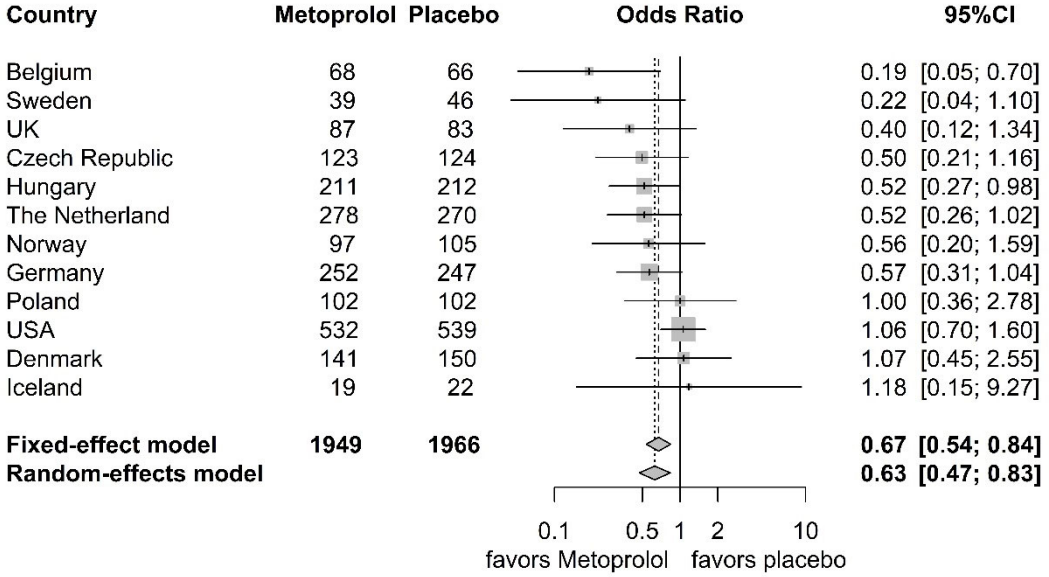

図 6 MERIT-HF 試験における国別オッズ比のフォレストプロット

治療効果と国の交互作用を評価するため，Cochran の Q 統計量を考えると，有意な交互作用は確認されず（p=0.18），Gail-Simon 検定でも有意な質的交互作用は確認されなかった（p=0.995）。Higgins のI^2統計量は 26.5%となり，低い地域間の異質性があると考えられたが，これまでの事例と同様に各国で十分な被験者数がないため，交互作用を特定するための検出力がないと考えられる。

続いて，厚生労働省の通知「国際共同治験における基本的考え方」（厚生労働省 2007）における方法 1 と方法 2 に基づく評価について考える。方法 1 において全体の治療効果$D_{all} = -\log(0.67)$とし，各国の治療効果$D_i = -\log(\hat{\theta}_i)$に対して，$D_i/D_{all} > \pi$を基準に治療効果が一貫しているかどうかを考える。$\pi = 0.5$とすると，ポーランド，アメリカ，デンマーク，アイスランドがこの基準を満たしていないため，全体集団の結果と一貫していない治療効果を持つ国となる。ここで，方法 1 に基づいて各国で必要な被験者数は全体の 22.4%である 894 名となり，アメリカのみがこれを満たしていたが，ポーランド，デンマーク，アイスランドは被験者数不足により評価不能となる。方法 2 においても，全体の治療効果$D_{all} = -\log(0.67)$とし，各国の治療効果$D_i =$

$-\log(\hat{\theta}_i)$に対して，全ての国の治療効果$D_1, D_2, \cdots, D_{12}$が 0 を上回るという基準を用いて一貫した結果であるかどうかを考えると，同様にポーランド，アメリカ，デンマーク，アイスランドが基準を満たさなかった。また，方法 2 に対する必要被験者数は全体の 15%以上と考えると，各国で 599 名が必要であり，アメリカ以外はこの被験者数を満たしていない。この場合，アメリカのみは方法 1 と方法 2 のいずれにおいても一貫していない地域として結論付ける事ができる。

これに対して提案法を適用する。ここでは，変量効果モデルにおいて地域間分散の DerSimonian-Laird 推定量$\hat{\tau}^2$が 0.0616 であるため，4 つの提案法で外れ値である地域と影響力のある地域の評価を行う。なお，ここでは共通して外れ値であると判断する基準を 5%とした。また，パラメトリックブートストラップ法におけるリサンプリングを 2400 回とした。最初に固定効果モデルにおける LOOCV 型のスチューデント化残差を用いた方法による結果を表 11 に示し，変量効果モデルにおける LOOCV 型のスチューデント化残差を表 12 に示した。アメリカのスチューデント化残差t_iは固定効果モデルで 2.579，変量効果モデルでは 2.578 となり，パラメトリックブートストラップ法によるt_iの標本分布における 97.5%点（固定効果モデルで 2.005，変量効果モデルで 2.393）をいずれも上回った。また，ベルギーのスチューデント化残差t_iは固定効果モデルで−1.942 となり，パラメトリックブートストラップ法によるt_iの標本分布における 2.5%点である−1.932 を下回った。

表 11 MERIT-HF 試験における固定効果モデルの LOOCV 型のスチューデント化残差を用いた外れ値となる国の評価結果

国	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
アメリカ	**2.579**	**−2.061**	**2.005**
ベルギー	**−1.942**	**−1.932**	**1.983**
スウェーデン	−1.373	−1.973	1.915
デンマーク	1.079	−1.994	2.030
イギリス	−0.868	−1.894	2.046
ハンガリー	−0.856	−1.955	1.922
オランダ	−0.796	−2.005	1.895
ポーランド	0.778	−1.978	1.974
チェコ共和国	−0.727	−2.045	1.940
ドイツ	−0.599	−1.952	1.978
アイスランド	0.533	−1.920	1.905
ノルウェー	−0.346	−1.888	1.952

表 12 MERIT-HF 試験における変量効果モデルの LOOCV 型のスチューデント化残差を用いた外れ値となる国の評価結果

国	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
アメリカ	**2.578**	**−2.426**	**2.393**
ベルギー	−1.769	−1.982	1.987
スウェーデン	−1.241	−2.054	2.024
デンマーク	1.082	−2.043	1.977
ポーランド	0.830	−1.921	2.074
イギリス	−0.697	−2.036	2.108
アイスランド	0.588	−1.960	1.967
チェコ共和国	−0.465	−2.077	2.016
ハンガリー	−0.456	−2.231	2.112
オランダ	−0.432	−2.010	2.089
ドイツ	−0.221	−2.176	2.026
ノルウェー	−0.175	−1.987	1.948

次に，固定効果モデル及び変量効果モデルにおけるモデルベースの尤度比検定に基づく方法による結果を

表 13，表 14 に示す。ここでもアメリカの尤度比検定統計量は固定効果モデルで 6.652，変量効果モデルで 5.532 となり，パラメトリックブートストラップ法による尤度比統計量の標本分布における 95%点（固定効果モデルで 3.889，変量効果モデルで 4.533）を大きく上回った。

表 13 MERIT-HF 試験における固定効果モデルの尤度比統計量を用いた外れ値となる国の評価結果

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
アメリカ	**6.652**	**3.889**	**0.011**
ベルギー	3.770	3.942	0.054
スウェーデン	1.886	3.704	0.174
デンマーク	1.164	3.738	0.287
ハンガリー	0.732	3.779	0.381
イギリス	0.753	3.946	0.401

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
ポーランド	0.605	3.786	0.426
オランダ	0.633	3.657	0.434
チェコ共和国	0.529	3.949	0.473
ドイツ	0.358	3.820	0.551
アイスランド	0.284	3.737	0.609
ノルウェー	0.120	3.639	0.728

表 14 MERIT-HF 試験における変量効果モデルの尤度比統計量を用いた外れ値となる国の評価結果

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
アメリカ	**5.532**	**4.533**	**0.027**
ベルギー	3.202	3.921	0.077
スウェーデン	1.596	3.835	0.208
デンマーク	1.211	3.817	0.278
ポーランド	0.693	4.025	0.418
イギリス	0.535	4.136	0.487
アイスランド	0.337	3.812	0.566
ハンガリー	0.314	4.157	0.612
チェコ共和国	0.278	4.017	0.615
オランダ	0.276	4.115	0.615
ドイツ	0.099	4.442	0.765
ノルウェー	0.044	3.962	0.836

最後に，全体分散の推定値及び地域間分散の推定値の相対的変化を評価した解析結果をそれぞれ表 15，表 16 に示す。ここでは，全体分散の相対的変化においては閾値となるパラメトリックブートストラップ法による$VRATIO$の標本分布の 5%点を下回っている国はなかったが，アメリカの$VRATIO$は最も低く，閾値にも近いため，影響力のある国であったと考えられる。また，地域間分散の相対的変化においてはアメリカがパラメトリックブートストラップ法による$TRATIO$の標本分布の 5%点を下回っていた。以上の結果より一貫してアメリカが外れ値となる治療効果を持つ国であり，影響力の

大きい国であった事がわかる。

表 15 MERIT-HF 試験における全体分散の相対的変化を評価した外れ値となる地域の評価結果

国	*VRATIO*	ブートストラップ 5%点
アメリカ	0.873	0.844
ベルギー	0.935	0.907
スウェーデン	0.991	0.928
アイスランド	1.043	0.953
イギリス	1.053	0.893
ノルウェー	1.124	0.884
ポーランド	1.127	0.863
デンマーク	1.131	0.835
チェコ共和国	1.140	0.828
オランダ	1.203	0.776
ハンガリー	1.213	0.780
ドイツ	1.274	0.763

表 16 MERIT-HF 試験における地域間分散の相対的変化を評価した外れ値となる地域の評価結果

国	*TRATIO*	ブートストラップ 5%点
アメリカ	**0.000**	**0.001**
ベルギー	0.710	0.576
スウェーデン	0.892	0.691
イギリス	1.007	0.517
アイスランド	1.070	0.788
デンマーク	1.109	0.261
チェコ共和国	1.124	0.216
ノルウェー	1.153	0.412
ポーランド	1.156	0.408
ハンガリー	1.187	0.003

国	$TRATIO$	ブートストラップ 5%点
オランダ	1.188	0.008
ドイツ	1.316	0.004

以上より，アメリカは 4 つの提案法で外れ値である可能性が示唆された。この結果は厚生労働省の通知に基づく方法による結果とも一致している。仮に何かの要因が治療効果に影響していると仮定し，主要な解析結果の感度解析としてアメリカを除外して解析したところ，固定効果モデルでオッズ比とその両側 95%信頼区間は 0.56 [0.42; 0.73]となった。この時，変量効果モデルにおいては地域間分散の DerSimonian-Laird 推定値$\hat{\tau}^2$が 0 となるため，固定効果モデルと同じ結果となった。これより，アメリカを除外しても有意な結果は変わらず，主要な解析結果の頑健性が示された。

MERIT-HF 試験に関して，Wedel らは地域間の違いやあらゆる背景因子の違いを議論したが，結論として以下の理由によりアメリカは偶然全体から外れた結果が得られたと考えられた（Wedel et al. 2001）。

- 治療効果と国による交互作用の解析を実施し，有意な結果は得られなかった
- アメリカの被験者において，重症度を示す New York Heart Association（NYHA）分類によって死亡割合が異なるが，生物学的な関連が不明であった
- もう一つの主要評価項目（全死亡及び理由によらない入院）と副次評価項目（全死亡及び心不全による入院）において，いずれもアメリカの結果がハザード比 1 を下回っていた
- 類薬においても有効性が示されている
- アメリカの部分集団解析において，被験者背景との交絡が特定できていない

そのため，本提案法のみで外れ値である事を決定する事はできず，更なる調査により偶然による結果である可能性も否定できない事は留意されたい。参考までに付録にもう一つの主要評価項目（全死亡及び理由によらない入院）と副次評価項目（全死亡及び心不全による入院）におけるフォレストプロット及び提案法を適用した結果をそれぞれ添付した。

3.6 シミュレーション実験

ICH E17 ガイドラインにおいて，各地域の被験者数配分について以下のような方法が紹介されている（ICH 2017）。

1. 比例配分：各地域の患者数や有病率に比例した被験者数配分
2. 均等配分：各地域への均等な被験者数配分
3. 効果の確保：全体集団の治療効果に対して特定の割合の治療効果が一つあるいは複数の地域で確保されるような被験者数配分
4. 地域の統計的有意性：各地域で統計学的に有意な結果が得られるような被験者数配分
5. 既定の最低被験者数：ある地域への既定の最低被験者数配分

これらの方法に対して，ガイドライン内では 3，4 は現実的ではなく，5 は科学的な妥当性がなければ勧められないと言及している。よって，1 又は 2 の方法が今後起こりえる被験者配分と考え，均等配分の場合と比例配分を想定して不均等配分の場合でシミュレーションを行った。

3.6.1. 均等配分の場合

提案法が潜在的に外れ値となる治療効果を持つ地域，または影響力のある地域を正しく検出できる事を確認するため，シミュレーションを行った。このシミュレーションでは$N(\mu_0, \tau^2)$からk個の地域の平均治療効果のデータをランダムに生成した。次に，$N(\mu_1, \tau^2)$からm個の地域の平均治療効果データを無作為に生成し，最初のm個の地域の平均治療効果データを新たに生成したデータに置き換えた。最後に，すべてのk個の地域に共通の地域内分散σ^2を付与した。3.5 節で検討した RECORD 試験と RENAAL 試験に基づき以下の 6 つのシナリオを想定した。

シナリオ 1：$k = 4$，$m = 1$，$\mu_0 = 1.0$，$\mu_1 = (0.5, 0.6, 0.7, 0.8)$，$\sigma^2 = 0.0225$及び$\tau^2 = 0.0025$のパラメータを持つ分布からデータを生成する。これは地域の数が少なく，地域間の異質性が極めて小さい場合を想定している。

シナリオ 2：k，m，μ_0，μ_1，σ^2はシナリオ 1 と同様のパラメータとし，$\tau^2 = 0.01$のパラメータを持つ分布からデータを生成する。これは，地域の数が少なく，地域間の異質性が小さい場合を想定している。

シナリオ 3：k，m，μ_0，μ_1，σ^2はシナリオ 1 と同様のパラメータとし，$\tau^2 = 0.04$パラメータを持つ分布からデータを生成する。これは地域の数が少なく，地域間の異質性が中程度である場合を想定している。

シナリオ 4：$k = 40$，$m = 3$，$\mu_0 = 0.5$，$\mu_1 = (2.2, 2.3, 2.4, 2.5)$，$\sigma^2 = 0.8$及び$\tau^2 = 0.0025$のパラメータを持つ分布からデータを生成する。これは地域の数が多く，地域間の異質性が極めて小さい場合を想定している。

シナリオ 5：k，m，μ_0，μ_1，σ^2はシナリオ 4 と同様のパラメータとし，$\tau^2 = 0.01$のパラメータを持つ分布からデータを生成する。これは地域の数が多く，地域間の異質性が小さい場合を想定している。

シナリオ 6：k，m，μ_0，μ_1，σ^2はシナリオ 4 と同様のパラメータとし，$\tau^2 = 0.04$のパラメータを持つ分布からデータを生成する。これは地域の数が多く，地域間の異質性が中程度である場合を想定している。

総被験者数は全体集団の治療効果により決定され，各地域の被験者数は総被験者数から割り当てられる。これに基づいて，地域の数が少ない場合には，各地域の被験者数が大きくなるために地域内分散は小さくなると想定した。一方，地域間分散は国際共同治験ごとに異なるため，地域間の異質性のパラメータτ^2を変更したいくつかのシナリオを用意した。提案した方法を上記の 6 つのシナリオで生成したデータに適用し，m個の地域を外れ値となる地域，又は影響力のある地域として検出できる割合を算出した。なお，外れ値となる地域を検出する基準を 5%に設定し，パラメトリックブート

ストラップ法による統計量の標本分布のパーセント点を閾値とした。なお，シミュレーション回数は 1000 回とした。

表 17 にシナリオ 1，2，3 の外れ値となる地域，又は影響力のある地域を検出する確率を示した。シナリオ 1 では全体集団と外れ値とした地域で治療効果に大きな差がある場合（$\mu_0 = 1.0$，$\mu_1 = 0.5$）に固定効果モデル及び変量効果モデルの LOOCV 型のスチューデント化残差に基づく方法で外れ値となる地域が正しく検出され（それぞれ 99.6%及び 93.1%），固定効果モデル及び変量効果モデルの尤度比検定に基づく方法でも外れ値となる地域が正しく検出された（それぞれ 99.6%及び 99.8%）。同じ治療効果に対し，全体分散の相対的変化に基づく方法では影響力のある地域を多く検出しなかったが（54.7%），各地域の分散の影響が地域間の異質性よりも相対的に大きいためと考えられる。一方，地域間分散の相対的変化に基づく方法では影響力のある地域を正確に検出した（100%）。治療効果の差がやや小さい場合（$\mu_0 = 1.0$，$\mu_1 = 0.6$）には変量効果モデルの LOOCV 型のスチューデント化残差及び全体分散の相対的変化に基づく方法は他の方法と比較して外れ値となる地域，又は影響力のある地域を検出する確率は大きく低下した。これより，これら 2 つの方法は他の方法よりも治療効果の差により影響を受ける事がわかる。全体集団と外れ値とした地域との間の治療効果の差が小さい場合（$\mu_0 = 1.0$，$\mu_1 = 0.7$，又は0.8），すべての提案法で外れ値とした地域と影響力のある地域をそれほど多く検出しなかった。シナリオ 2 では全ての提案法で外れ値となる地域，又は影響力のある地域を検出する確率はシナリオ 1 より低かった。特に，変量効果モデルに対する LOOCV 型のスチューデント化残差に基づく方法は他の方法に比べて外れ値となる地域を検出しなかった。これは変量効果モデルに対する LOOCV 型のスチューデント化残差に基づく方法が他の方法よりも地域間の異質性により大きな影響を持つ事を意味している。シナリオ 3 では全ての方法で外れ値となる地域，又は影響力のある地域を検出する確率はシナリオ 2 の結果より低かった。従って，この結果は地域の数が少ない場合に地域間の異質性のパラメータτ^2が減少するにつれて，すべての提案法で外れ値となる地域，又は影響力のある地域を検出する確率を低下させる事を示した。地域間の異質性がより大きい場合（シナリオ 2 及び 3），変量効果モデルに対する LOOCV 型のスチューデント化残差及び尤度比検定に基づく方法において外れ値となる地域を検出する確率は固定効果モデルよりも低くなった。これは地域間の異質性を考慮する事により，変量効果モデルの全体分散が固定効果モデルの全体分散よりも大きくなるためであり，この時，変量効果モデルの治療効果の差は固定効果モデルの場合より小さく評価される。しかしながら，変量効果モデルは外れ値となる地域の検出だけでなく，全体分散の相対的変化を評価する方法や地域間分散の相対的変化を評価する方法を用いた影響力の診断を可能にするため，変量効果モデルは固定効果モデルよりも包括的に評価が可能となる。さらに，地域間の異質性が小さい場合（シナリオ 1），治療効果の差が大きい場合（$\mu_0 = 1.0$，$\mu_1 = 0.5$）と治療効果の差が小さい場合（$\mu_0 = 1.0$，$\mu_1 = 0.8$）の外れ値となる地域を検出する確率の差は大きい。一方，地域間に中程度の異質性（シナリオ 3）が存在する場合は治療効果の差が大きい場合と治療効果が小さい場合の外れ値となる地域を検出する確率の差は減少する。例えば，固定効果モデルにおいて LOOCV 型のスチューデント化残差に基づく方法で考えると，シナリオ 1 の治療効果の差が大きい場合の外れ値となる地域を検出する確率（99.6%）と治療効果の差が小さい場合の外れ値となる地域を検出する確率（0.

3%）の差は 99.3%となる。一方，シナリオ 3 の治療効果の差が大きい場合の外れ値となる地域を検出する確率（74.3%）と治療効果の差が小さい場合の外れ値となる地域を検出する確率（26.9%）の差は 47.4%となる。地域間の異質性が大きい場合には治療効果の差は比較的小さく評価されるため，LOOCV 型のスチューデント化残差及び尤度比統計量に基づく方法では治療効果の差の変化が外れ値となる地域を検出する確率に及ぼす影響は大きくない。全体分散の相対的変化に基づく方法及び地域間分散の相対的変化に基づく方法について，地域間の異質性が大きい場合には，外れ値とした地域とそれ以外の地域との差を相対的に小さく評価する。従って，これらの方法でも同様に地域間の異質性に応じて外れ値となる地域を検出する確率は変動する。

表 17 シナリオ 1，2，3（均等配分）に対する外れ値となる地域を検出，又は影響力のある地域を特定する確率

		スチューデント化残差		尤度比統計量		全体分散の相対的変化	地域間分散の相対的変化
シナリオ	治療効果 $\mu_0{:}\mu_1$	固定効果モデル	変量効果モデル	固定効果モデル	変量効果モデル		
1	1.0:0.5	99.6%	93.1%	99.6%	99.8%	54.7%	100.0%
	1.0:0.6	84.7%	49.4%	84.5%	89.7%	6.0%	97.1%
	1.0:0.7	24.1%	8.9%	23.3%	33.0%	0.1%	56.0%
	1.0:0.8	0.3%	0.3%	0.3%	1.0%	0.0%	5.1%
2	1.0:0.5	91.1%	66.8%	90.7%	83.7%	52.6%	88.7%
	1.0:0.6	69.8%	38.1%	70.2%	61.5%	23.1%	79.5%
	1.0:0.7	36.0%	15.4%	35.7%	32.9%	5.7%	54.5%
	1.0:0.8	11.5%	2.8%	11.3%	10.2%	0.6%	26.2%
3	1.0:0.5	74.3%	32.3%	74.4%	39.2%	30.5%	36.7%
	1.0:0.6	58.7%	22.5%	59.7%	27.5%	19.7%	31.9%
	1.0:0.7	42.4%	13.5%	42.9%	18.2%	12.0%	25.2%
	1.0:0.8	26.9%	7.5%	27.6%	10.8%	6.0%	19.1%

表 18 にシナリオ 4, 5, 6 の外れ値となる地域を検出する確率を示した。これらのシナリオでは地域間分散を推定する際にほとんどの場合で$\hat{\tau}^2 = 0$となるため，固定効果モデルに対する LOOCV 型のスチューデント化残差と尤度比検定に基づく方法のみによるシミュレーションを行った。さらに，外れ値となる地域として 3 地域の治療効果をその他の地域の治療効果と異なる値とし，そのうちいくつの外れ値となる地域を検出できるかを確認した。シナリオ 4 では，外れ値となる地域とその他の地域の間で治療効果が大きく異なる場合に（$\mu_0 = 0.5$，$\mu_1 = 2.5$），LOOCV 型のスチューデント化残差及び尤度比検定に基づく方法はいずれも少なくとも一つの外れ値となる地域を検出し（いずれも 100%），3 つの外れ値となる地域についても高い確率ですべてを正確に

検出した（それぞれ 90.2%及び 93.6%）。治療効果の差が減少するにつれて，いずれの方法でも 3 つの外れ値となる地域すべてを検出する確率は低下した。シナリオ 5 と 6 では，治療効果が大きく異なる場合に（$\mu_0 = 0.5$，$\mu_1 = 2.5$），LOOCV 型のスチューデント化残差と尤度比検定に基づく方法の両方で少なくとも 1 つの外れ値となる地域を正しく検出する確率は高かったが 3 つ全ての外れ値となる地域を検出する確率は地域間の異質性が大きくなるにつれて，減少した。また，治療効果の差が減少するにつれて，外れ値となる地域すべてを検出する確率は低下した。

表 18 シナリオ 4，5，6（均等配分）に対する外れ値となる地域を検出，又は影響力のある地域を特定する確率

シナリオ	治療効果 μ_0:μ_1	スチューデント化残差				尤度比統計量			
		0	1	2	3	0	1	2	3
4	0.5:2.5	0.0%	0.1%	9.7%	90.2%	0.0%	0.2%	6.2%	93.6%
	0.5:2.4	3.5%	22.8%	43.1%	30.6%	3.1%	19.1%	46.2%	31.6%
	0.5:2.3	53.8%	35.4%	9.9%	0.9%	62.6%	31.5%	5.7%	0.2%
	0.5:2.2	96.0%	4.0%	0.0%	0.0%	98.0%	2.0%	0.0%	0.0%
5	0.5:2.5	0.0%	3.7%	29.9%	66.4%	0.2%	3.2%	29.5%	67.1%
	0.5:2.4	6.2%	28.4%	43.6%	21.8%	6.0%	27.3%	45.8%	20.9%
	0.5:2.3	35.3%	43.6%	18.0%	3.1%	38.8%	44.1%	18.0%	1.4%
	0.5:2.2	75.6%	22.9%	1.5%	0.0%	80.0%	18.8%	1.2%	0.0%
6	0.5:2.5	2.1%	15.5%	45.4%	37.0%	1.8%	14.9%	45.4%	37.9%
	0.5:2.4	9.3%	34.4%	39.2%	17.1%	8.9%	32.6%	42.0%	16.5%
	0.5:2.3	23.3%	45.3%	25.6%	5.8%	23.7%	46.3%	24.9%	5.1%
	0.5:2.2	44.8%	42.5%	11.6%	1.1%	47.4%	41.0%	10.8%	0.8%

3.6.2. 不均等配分の場合

前節で定義したシナリオにおいて，実際は各地域の患者数によって被験者数が異なる場合が多い。そのため，シナリオ 1～3 においては，治療効果を$\mu_0 = 1.0$と$\mu_1 = 0.5$に固定した場合に外れ値となる地域の地域内分散σ^2を 0.04，0.0625，0.09 に変化させてシミュレーションを行った。これは外れ値である地域の分散を大きくしているため，外れ値となる地域のみの被験者数が少ない場合を想定したものである。その結果を表 19 に示した。全体的に地域内分散が大きくなるにつれて外れ値を検出する確率は低下したが，地域間分散の相対的変化に基づく方法のみ，地域内分散の影響を受け難かった。これは統計量に地域内分散を含んでいないため，影響を受け難いと考えられる。また，全体分散の相対的変化に基づく方法に関しては大きく影響を受け，外れ値となる地域を検出する確率を低下させた。これは直接地域内分散が統計量に影響するため

と考えられる。LOOCV 型のスチューデント化残差及び尤度比検定に基づく方法においては同程度の影響を受けた。この結果より，極端に外れ値となる地域の被験者数が小さい場合には外れ値となる地域を検出する確率は低くなるが，ある程度均等割付した場合の被験者数に近ければ外れ値となる地域を検出する確率は維持できると考えられる。

表 19 シナリオ 1, 2, 3（不均等配分）に対する外れ値となる地域を検出，又は影響力のある地域を特定する確率

		スチューデント化残差		尤度比統計量		全体分散の相対的変化	地域間分散の相対的変化
シナリオ	地域内分散*	固定効果モデル	変量効果モデル	固定効果モデル	変量効果モデル		
1	0.04	87.7%	91.8%	88.2%	92.9%	20.3%	100%
	0.0625	37.6%	50.3%	36.1%	47.3%	0.8%	99.4%
	0.09	2.6%	6.0%	2.0%	3.6%	0%	97.1%
2	0.04	72.2%	60.3%	72.3%	71.6%	37.9%	87.1%
	0.0625	41.8%	41.4%	42.2%	45.3%	17.9%	83.3%
	0.09	16.4%	18.7%	15.5%	18.2%	3.7%	79.5%
3	0.04	65.6%	42.5%	66.1%	51.6%	35.6%	56.4%
	0.0625	44.4%	31.9%	44.8%	39.3%	22.7%	53.8%
	0.09	24.6%	20.1%	24.4%	22.6%	11.9%	53.9%

* 外れ値となる地域の地域内分散

また，シナリオ 4～6 においては，治療効果を$\mu_0 = 0.5$と$\mu_1 = 2.5$に固定した場合に 3 つの外れ値となる地域に対して，地域内分散σ^2が 1.0 となる地域数を 1 地域ずつ増やしてシミュレーションを行った。その結果を表 20 に示した。LOOCV 型のスチューデント化残差及び尤度比検定に基づく方法のいずれにおいても分散が大きくなった地域は外れ値として検出され難くなり，同程度の影響が見られた。しかしながら，地域数が 40 であり，もともとの各地域の被験者数は小さい事が想定されるため，これは外れ値となる地域の被験者数が極めて小さい場合であることが想定される。

表 20 シナリオ 4，5，6（不均等配分）に対する外れ値となる地域を検出する確率

シナリオ	地域内分散*	スチューデント化残差				尤度比統計量			
		0	1	2	3	0	1	2	3
4	1.0:0.8:0.8	0%	4.0%	80.8%	15.2%	0.1%	2.9%	86.4%	10.6%
	1.0:1.0:0.8	1.2%	64.1%	30.3%	4.4%	0.6%	73.7%	24.3%	1.4%
	1.0:1.0:1.0	46.9%	40.7%	10.6%	1.8%	57.5%	35.0%	7.7%	0.2%

シナリオ	地域内分散*	スチューデント化残差				尤度比統計量			
		0	1	2	3	0	1	2	3
5	1.0:0.8:0.8	0.9%	13.5%	64.6%	21.0%	0.6%	15.3%	65.9%	18.2%
	1.0:1.0:0.8	5.4%	48.0%	39.3%	7.3%	5.4%	53.0%	36.0%	5.6%
	1.0:1.0:1.0	32.3%	44.4%	19.6%	3.7%	37.0%	45.1%	15.4%	2.5%
6	1.0:0.8:0.8	4.0%	27.0%	50.6%	18.4%	4.0%	26.8%	51.3%	17.9%
	1.0:1.0:0.8	9.9%	40.2%	39.7%	10.2%	10.1%	41.2%	40.0%	8.7%
	1.0:1.0:1.0	22.8%	44.8%	25.6%	6.8%	23.0%	46.3%	25.5%	5.2%

* 外れ値となる 3 地域の地域内分散

3.6.3. 検定の不偏性について

尤度比検定の不偏性を検証するため，治療効果を$\mu_0 = 0.99$と$\mu_1 = 1.00$に固定し，対立仮説$H_1{:}\zeta = 0.01$の下で地域数，地域内分散，地域間分散を変化させ，異なる治療効果を持つ 1 地域に対して，該当地域が異なる治療効果を持つ地域として検出される確率をシミュレーションで算出した（表 21，表 22）。なお，有意水準を 5%に設定し，パラメトリックブートストラップ法による統計量の標本分布のパーセント点を閾値とした。この時，地域数が少ない場合（$k = 4$）においても，地域内分散σ^2が 0.0225 以下，かつ地域間分散τ^2が 0.04 以上であれば，不偏性を保つ事が可能である事が示された。また，地域数が多くなるにつれて，検出確率は有意水準である 5%に近づいていく事も示された。

表 21 尤度比検定において地域数と地域間分散を変化させた場合の異なる治療効果を持つ地域を検出する確率

治療効果 $\mu_0{:}\mu_1$	地域内分散	地域数	地域間分散	検出確率
0.99:1.00	0.0225	4	0.04	6.6%
			0.09	9.3%
			0.16	9.5%
			0.25	8.9%
		8	0.04	6.6%
			0.09	7.0%
			0.16	6.5%
			0.25	6.1%
		16	0.04	6.6%
			0.09	5.8%
			0.16	5.3%

治療効果 μ_0:μ_1	地域内分散	地域数	地域間分散	検出確率
			0.25	5.6%
		32	0.04	5.1%
			0.09	5.3%
			0.16	5.3%
			0.25	5.0%

表 22 尤度比検定において地域数と地域内分散を変化させた場合の異なる治療効果を持つ地域を検出する確率

治療効果 μ_0:μ_1	地域間分散	地域数	地域内分散	検出確率
0.99:1.00	0.04	4	0.0025	8.2%
			0.01	9.1%
			0.0225	6.6%
		8	0.0025	5.5%
			0.01	6.5%
			0.0225	6.6%
		16	0.0025	5.5%
			0.01	5.9%
			0.0225	6.6%
		32	0.0025	5.0%
			0.01	5.4%
			0.0225	5.1%

3.7 考察

3.7.1. 提案法の適用

本研究では国際共同治験の枠組みで外れ値となる地域の検出と影響力の評価を目的に 4 つの提案法を提案した。4 つの提案法はそれぞれ異なる特徴を持ち，より頑健な結論を導くためにはすべての解析が実施される事が望まれる。シミュレーション実験の結果より LOOCV 型のスチューデント化残差と尤度比検定に基づく方法では概ね同様の結果を与えるが，LOOCV 型のスチューデント化残差に基づく方法の方が地域間の異質性の影響は大きく，保守的に外れ値を検出する傾向にある。そのため，推定された地域間の異質性の確からしさによって，使い分けることは可能であると考える。

また，LOOCV 型のスチューデント化残差に基づく方法と尤度比検定に基づく方法では外れ値となる地域とそれ以外の地域の平均パラメータの乖離を評価していたが，全体分散の相対的変化に基づく方法と地域間分散の相対的変化に基づく方法を用いて，各地域の治療効果の分散を評価する事も有用である。この二つの方法においても異なる観点での評価となり，地域間分散τ^2と地域内分散σ_i^2の大きさによっても使い分ける必要がある。地域間分散τ^2が地域内分散σ_i^2に比べて小さい時には全体分散の相対的変化は小さくなり，単に各地域の被験者規模を反映したものとなる可能性がある。よって，地域間分散τ^2の相対的変化も併せて確認する必要がある。また，分散に基づいた方法では外れ値の基準を超えていなかったとしても，どの程度の影響があったのかを評価する指標となりえる。統計量が 1 を下回る場合に影響があると考えられ，その影響度もどの程度 0 に近いかで参考値として用いる事が可能となる。これにより，潜在的な外れ値となる地域に対して無視できるほどの影響力であれば，試験全体の結果は頑健であると考えられ，潜在的な外れ値となる地域に対して無視できないほどの大きな影響力があれば，全地域共通の治療効果として得られた結果が全ての地域に一般化できない可能性が考えられる。それ故，本論文では提案法のいずれか一つを推奨するものではなく，あらゆる角度から外れ値となる地域の可能性を評価するうえで全ての提案法を適用する事を推奨する。また，各地域の推定値$\hat{\mu}$，地域内分散σ_i^2，及び地域間分散の推定値$\hat{\tau}^2$のバランスを考慮し，外れ値となる地域の可能性を評価する必要がある。さらに，メトプロロールの事例の通り，提案法で外れ値となる地域として検出されたとしても，周辺情報から外れ値となる地域ではないという結論が妥当な場合もあり得る。そのため，提案法により外れ値である事が疑われる地域を検出し，検出された地域に対して治療効果に関連する要因の探索などの更なる調査を行う事を推奨する。また，シミュレーションの結果より 4 つの提案法は被験者数が少ない場合には外れ値を検出する確率が低下するが，極端に被験者数が少なくなければ評価が可能であると考えられる。事例においても RECORD 試験ではスリランカの被験者数は 49 名であり，全体の 0.6%程度の被験者規模であるものの，外れ値として検出できている。さらに，RENAAL 試験ではイスラエルの被験者数は 37 名であり，全体の 2.5%程度であるが，外れ値として検出できている。そのため，従来法より被験者数を必要とせずに一貫性の評価が可能であると考えられる。一方，一つの地域の被験者数が極端に少ない場合は，地域を併合する事を考える必要がある。その場合はデータを見てから恣意的に地域を分類するのではなく，事前に特定されている類似した背景情報を持つ地域や地理的に近い地域のような分類を予め定義し，併合した地域を一つの地域として本提案法を適用すべきであると考える。ただし，提案法は偶然の誤差を前提に外れ値となる地域を検出する事となり，例えば，5%の有意水準で尤度比検定に基づく方法で外れ値となる地域を検出する場合，外れ値となる地域は全地域数の 5%の地域数で検出されてしまうことに留意されたい。さらに，本提案法では検定の多重性については考慮されていないが，基本的には，本提案法は検証的な解析ではなく，影響力評価という位置づけでの解析になるため，厳密な調整は不要であると考える。しかしながら，必要であれば，第一種の過誤確率の調整ではなく，第二種の過誤確率を調整する事は有用であると考える。これについては今後の検討課題としたい。なお，潜在的な外れ値として検出された地域に対して，偶然によって治療効果に影響を与える可能性のある内因性，外因性の要因が検出されてしまう事を避けるため，要因探索のために本提案

法を繰り返し使用する事は想定していない。本提案法により潜在的な外れ値となる地域を検出し，潜在的な外れ値となる地域に対して多地域と比べて臨床的に重要な差異がない場合には全体の結果を各地域の結果に外挿する事が可能であると考えられる。一方，潜在的な外れ値となる地域に対して，臨床的に重要な差異がある場合には，治療効果に影響を与える内因性，外因性の要因を特定する事で，各規制当局の意思決定に役立つ情報を提供できると考える。特に，潜在的な外れ値となる地域に対して，地域間の治療効果に影響を与える要因の分布の差異により見かけの差が生じている可能性を検討する事は有意義であると考える。

3.7.2. 提案法の位置づけ

本提案法は探索的な位置付けであるが，ある程度の基準を設けて解析を実施するため，検証的な意味合いも含んでいる。ただし，本提案法により潜在的な外れ値となる地域として検出された地域がそのまま治療効果の高い地域，又は治療効果の低い地域と結論づける事はできない。偶然によって検出されている可能性もあるため，潜在的な外れ値となる地域に対して，治療効果に影響を与える内因性，外因性の要因の偏り等を調査し，地域間の治療効果の差異が無視できる程度のものであるかどうかも鑑みて，新薬の有用性を判断すべきである。なお，本提案法は探索的な位置付けで適用されるため，本提案法を適用する事で積極的に多くの外れ値となる地域の候補を挙げ，更なる調査によって治療効果に関連する要因を評価する方が保守的であると考える。そのため，本提案法では第二種の過誤確率を重視し，いくつかの状況を想定した上で第二種の過誤がどの程度発生する可能性があるかを評価する事を推奨する。

また，本提案法によって検出された潜在的な外れ値となる地域に対して，該当地域を除外した解析は主解析の結果の頑健性を示す感度分析の位置づけで実施され，感度分析の結果のみから該当地域以外の地域の治療効果が高い，又は治療効果が低いという結論に至るものではない。なお，外れ値が主解析の結果に過度に影響を及ぼしている場合には，潜在的な外れ値となる地域を除外するのではなく，全体集団の効果の推定値と個々の地域のデータを利用した推定値の加重平均を用いる方法（縮小推定量）も有用である（ICH 2017, Quan et al. 2013）。

3.7.3. パラメトリックブートストラップ法

最近ではいくつかの研究で，小標本下での尤度比検定において，パラメトリックブートストラップによる調整が推論の妥当性を改善できる事が示されており（Noma et al. 2018, Stein et al. 2014, Ukyo et al. 2019），本研究でも小標本下での提案法の統計量のばらつきを適切に考慮するためにパラメトリックブートストラップ法を適用した。最近の計算機の発展により，ブートストラップの計算時間は極端に短くなり，ブートストラップ法の使用が推奨されてきている。一方，ブートストラップ法では帰無仮説が正しいと仮定して，推定されたパラメータから帰無分布を推定しているが，母集団の真の分布は帰無仮説で想定する母集団とは異なっている可能性があるため，そこから抽出された初期標本に基づいて得られるブートストラップ近似分布は提案した統計量の帰無分布の合理的な近似とは考えられない可能性が懸念される（汪と田栗 1996）。

3.7.4. マスキング効果とスワンピング効果

一般に外れ値が複数存在する場合，LOOCV 型の外れ値の検出法においては外れ値同士が影響し合ってしまい，正しく外れ値が検出されない場合がある（Acuna and Rodriguez 2004）。これはマスキング効果，又はスワンピング効果と呼ばれている。マスキング効果は 2 つの外れ値が存在する場合に 2 つ目の外れ値が 1 つ目の外れ値を覆い隠してしまう場合である。つまり，マスキング効果は外れ値のグループが分布を歪ませ，平均からの外れ値の距離が小さくなってしまう場合に生じる。この場合に外れ値を 1 つ取り除いた時に別の外れ値が現れるような事がある。これは外れ値を外れ値と検出する事を妨げるため，第二種の過誤確率を増大させている。一方，スワンピング効果は外れ値が存在する場合にのみある観測値が外れ値と考えられる場合である。つまり，スワンピング効果は外れ値のグループが分布を歪ませ，平均からの距離が大きくなってしまう場合に生じる。この場合，外れ値がある事によって他の正常値を外れ値としてしまっている。これは正常値を外れ値としてしまうため，第一種の過誤確率を増大させている。マスキング効果やスワンピング効果においては，パラメータの推定にロバスト推定を用いる事も有用であると考えられる。または，ロサルタンの事例のように同様の背景因子により異なる治療効果をもつ可能性のある地域を併合した後に外れ値であるかどうかを確認するという事も回避する一つの方法となりえるかもしれない。

4. 結論

医薬品開発の国際化により国際共同治験の実施は臨床研究における標準的な戦略になりつつある。しかし，国際共同治験では様々な治療に関連する要因の分布が地域間で異なる事により，治療効果が各地域で異なる場合がある。国際共同治験の主目的は地域共通の治療効果を検証することであるが，この地域間の治療効果の違いが主目的に対する結論に影響を及ぼす可能性がある。そこで，地域共通の治療効果において有効性が示せた場合に各地域で同様の治療効果があるかどうかの一貫性を評価し，一貫していない場合にその影響を評価する事は国際共同治験を実施する上で重要な統計的課題となっている。

この課題に対して，これまで，フォレストプロットによって視覚的に治療効果が一貫していない地域があるかどうかが確認されてきた。しかしながら，フォレストプロットによる評価では多くの場合で各地域の治療効果の推定値の信頼区間が重なり合い，治療効果が一貫していない地域を判断する事は困難である。また，地域と治療効果の交互作用を評価する検定が提案されているが，国際共同治験の主目的は地域共通の治療効果の検証であるため，交互作用の検定を行うには十分な被験者数が確保されておらず，十分な検出力がない。さらに，最も頻繁に用いられる厚生労働省の通知（厚生労働省 2007）に示されている方法では各地域の治療効果の推定値が事前に定めた基準を超えるかどうかにより治療効果の地域間の一貫性を評価しており，各地域の真の治療効果が一貫していると仮定した場合に高い確率で基準を上回るような被験者数が要求されている。この時，国際共同治験への参加地域数が多い場合に各地域で要求され

ている被験者数を確保できず，結論に至る事ができない場合がある。そこで本研究では異なる観点から 4 つの方法を提案した。これらの 4 つの方法は明確な基準を設定する事で客観的に外れ値となる地域を検出でき，分散の相対的変化を評価する方法では各地域に対する統計量が地域共通の治療効果への影響力の指標となる。さらに，従来法のように治療効果の推定値のみを用いて評価を行うのではなく，地域内の分散や地域間の異質性も評価することでより詳細に一貫性の評価が可能となる。

本研究では，国際共同治験の治療効果の地域間の一貫性評価の実践的な有用性についても評価した。まず，国際共同治験では事例解析のように地域数が増加するにつれて，各地域の被験者数が減少するという性質があり，このような性質の下でも十分に外れ値を検出できることが示された。さらに，事例解析とシミュレーションの結果から各地域の被験者数が極端に小さくなければ，提案法による評価が可能であると考えられた。ここでは地域間差に焦点を当てたが，当然一つの地域の中でも異なる特性を持つ被験者が存在する可能性はあり，それにより異なる治療効果を持つ場合がある。日米 EU 医薬品規制調和国際会議ガイドラインの E17（International Conference ICH 2017）では，地域は内因性要因，外因性要因等の地域差の原因となる未知の要因の指標とされている。従って，地域間の異質性が予期せず観察された場合，結果の解釈を助けるために地域間のばらつきに寄与する可能性のある内因性及び外因性因子のさらなる調査が必要である。いくつかの様々な治療に関連する因子が特定された場合，予測因子のハット行列から推定されるてこ比のような予測ベースの異常値検出法を適用できるが，地域間で被験者数が異なる中で複数の調整変数を含むモデルは解釈を複雑にする可能性がある。そのため，ここでは潜在的な外れ値となる地域，又は影響力のある地域を評価するための単純な枠組みとして，外れ値か否かを評価する外れ値検出法を与えた。外れ値か否かの評価に基づく方法で外れ値となる地域，又は影響力のある地域を検出する場合，外れ値となる地域，又は影響力のある地域において治療に関連する可能性のある因子の分布を検討し，治療効果と関連した因子を特定する事が重要である。理想的には，様々な関連因子が外れ値となる地域と影響力のある地域を検出するために考慮され，同時に関連する因子を調整する方法によって評価されるべきであるが，それは今後の検討課題であると考える。

結論として，本研究では国際共同治験の治療効果の地域間の一貫性評価において，外れ値となる地域，又は影響力のある地域を評価するための効果的なツールを提案した。実データへの応用で示されたように，様々な理由により外れ値となる治療効果を持つ地域や影響力のある地域を持つ国際共同治験は存在し，これに対して主目的である地域共通の治療効果への影響に関して結論付けられない事は避けられるべきである。そこで，今後の国際共同治験での一貫性の評価において提案法が有用なツールとして利用でき，地域共通の治療効果に対するエビデンス構築に有用となる事が期待される。

5. 謝辞

博士課程において，主任指導教員として終始適切な助言を賜り，丁寧に指導して下さった野間久史准教授に，深くお礼申し上げます。特に，研究で行き詰った際には大変貴重なご意見を頂き，前に進むことができました。また，指導のために多くの時間を割いて頂き，多くの事を学ばせて頂きました。学位審査におきましては，ご多忙に

も関わらず，博士論文の審査委員をお引き受け下さりました二宮嘉行教授，逸見昌之准教授，京都大学の土居正明准教授に謹んで感謝申し上げます。審査の過程で多くの有益なご指導を頂きました。また，逸見昌之准教授には副指導教員もご担当頂きました。重ねて感謝申し上げます。本研究の投稿論文に関しましては，筑波大学の五所正彦教授には共著者として，大変有益なご助言，ご指導を頂きました。厚くお礼申し上げます。また，三年間の学生生活の中で充実した研究環境とサポートを提供して下さった総合研究大学院大学および統計数理研究所の皆様にお礼申し上げます。ノバルティス ファーマ株式会社の皆様にはいろいろと業務の上でご配慮頂き，かつ一部の方には相談に乗って頂きました。私一人の力では博士課程に進むという決断や，問題解決に至る事ができなかったと思います。心より感謝申し上げます。最後に，常に励まし，寄り添い，支えて頂いた私の家族に感謝致します。

6. 参考文献

1. Acuna, E., and Rodriguez, C.A. (2004), "A Meta analysis study of outlier detection methods in classification." *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*:1-25.
2. Asano, K., Tanaka, A., Sato, T., and Uyama, Y. (2013), "Regulatory challenges in the review of data from global clinical trials: the PMDA perspective." *Clin Pharmacol Ther*, 94:195-8.
3. Belsley, D.A, Kuh, E, and Welsch, R.E. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*. Edited by John Wiley & Sons.
4. Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2010), "A basic introduction to fixed-effect and random-effects models for meta-analysis." *Res Synth Methods*, 1:97-111.
5. Brenner, B. M., Cooper, M. E., de Zeeuw, D., Keane, W. F., Mitch, W. E., Parving, H. H., Remuzzi, G., Snapinn, S. M., Zhang, Z., Shahinfar, S., and Investigators, Renaal Study. (2001), "Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy." *N Engl J Med*, 345:861-9.
6. Chen, J., Quan, H., Binkowitz, B., Ouyang, S. P., Tanaka, Y., Li, G., Menjoge, S., Ibia, E., and Consistency Workstream of the Ph, R. M. A. Mrct Key Issue Team. (2010), "Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review." *Pharm Stat*, 9:242-53.
7. Chen, Joshua, Quan, Hui, Gallo, Paul, Menjoge, Shailendra, Luo, Xiaolong, Tanaka, Yoko, Li, Gang, Ouyang, S. Peter, Binkowitz, Bruce, Ibia, Ekopimo, Talerico, Steven, and Ikeda, Kimitoshi. (2011), "Consistency of Treatment Effect across Regions in Multiregional Clinical Trials, Part 1: Design Considerations." *Drug Information Journal*, 45:595-602.
8. Chen, X., Lu, N., Nair, R., Xu, Y., Kang, C., Huang, Q., Li, N., and Chen, H. (2012), "Decision rules and associated sample size planning for regional approval utilizing multiregional clinical trials." *J Biopharm Stat*, 22:1001-18.
9. DerSimonian, R., and Laird, N. (1986), "Meta-analysis in clinical trials." *Control Clin Trials.*, 7:177-188.
10. Diao, G., Zeng, D., Ibrahim, J. G., Rong, A., Lee, O., Zhang, K., and Chen, Q. (2017), "Statistical design of noninferiority multiple region clinical trials to assess global and

consistent treatment effects." *J Biopharm Stat*, 27:933-944.

11. Efron, Bradley, and Tibshirani, Robert. 1994. *An introduction to the bootstrap*. New York: Chapman & Hall.
12. FDA. (2009), "Cardiovascular and Renal Drugs Advisory Committees and Meeting Materials.".
13. Gail, M., and Simon, R. (1985), "Testing for Qualitative Interactions between Treatment Effects and Patient Subsets." *Biometrics*, 41.
14. Group, MERIT-HF Study. (1999), "Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trial in-Congestive Heart Failure (MERIT-HF)." *The Lancet*, 353:2001-2007.
15. Guo, H., Chen, J., and Quan, H. (2016), "Evaluation of local treatment effect by borrowing information from similar countries in multi-regional clinical trials." *Stat Med*, 35:671-84.
16. Hedges, LV., and Olkin, I. 1985. *Statistical Methods for Meta-Analysis*.
17. Higgins, J. P., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003), "Measuring inconsistency in meta-analyses." *BMJ*, 327:557-60.
18. Higgins, Julian P. T., Thomas, James, Chandler, Jacqueline, Cumpston, Miranda, Li, Tianjing, Page, Matthew J., and Welch, Vivian A. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*.
19. Hjalmarson, A., Goldstein, S., Fagerberg, B., Wedel, H., Waagstein, F., Kjekshus, J., Wikstrand, J., El Allaf, D., Vitovec, J., Aldershvile, J., Halinen, M., Dietz, R., Neuhaus, K. L., Janosi, A., Thorgeirsson, G., Dunselman, P. H., Gullestad, L., Kuch, J., Herlitz, J., Rickenbacher, P., Ball, S., Gottlieb, S., and Deedwania, P. (2000), "Effects of controlled-release metoprolol on total mortality, hospitalizations, and well-being in patients with heart failure: the Metoprolol CR/XL Randomized Intervention Trial in congestive heart failure (MERIT-HF). MERIT-HF Study Group." *JAMA*, 283:1295-302.
20. Hung, H. M., Wang, S. J., and O'Neill, R. T. (2010), "Consideration of regional difference in design and analysis of multi-regional trials." *Pharm Stat*, 9:173-8.
21. ICH. (1998), "Ethnic factors in the acceptability of foreign clinical data."
22. ICH. (2017), "General Principles for Planning and Design of Multi-Regional Clinical Trials.".
23. Ichimaru, K., Toyoshima, S., and Uyama, Y. (2010), "Effective global drug development

strategy for obtaining regulatory approval in Japan in the context of ethnicity-related drug response factors." *Clin Pharmacol Ther*, 87:362-6.

24. Ikeda, K., and Bretz, F. (2010), "Sample size and proportion of Japanese patients in multi-regional trials." *Pharm Stat*, 9:207-16.
25. Kawai, N., C., Chuang-Stein, Komiyama, O., and Li, Y. (2008), "An approach to rationalize partitioning sample size into individual regions in a multiregional trial." *Drug Information Journal*, 42:139-147.
26. Kim, Saemina, and Kang, Seung-Ho. (2019), "Hierarchical Linear Models for Multiregional Clinical Trials." *Statistics in Biopharmaceutical Research*, 12:334-343.
27. Knapp, G., and Hartung, J. (2003), "Improved tests for a random effects meta-regression with a single covariate." *Stat Med*, 22:2693-710.
28. Ko, F. S., Tsou, H. H., Liu, J. P., and Hsiao, C. F. (2010), "Sample size determination for a specific region in a multiregional trial." *J Biopharm Stat*, 20:870-85.
29. Koshimizu, Takashi. (2003), "Statistical Considerations on Bridging Strategy through Joining a Multi-regional Study." *Japanese Journal of Biometrics*, 24:S99-S104.
30. Lin, D. Y., and Zeng, D. (2010), "On the relative efficiency of using summary statistics versus individual-level data in meta-analysis." *Biometrika*, 97:321-332.
31. Liu, J. T., Tsou, H. H., Gordon Lan, K. K., Chen, C. T., Lai, Y. H., Chang, W. J., Tzeng, C. S., and Hsiao, C. F. (2016), "Assessing the consistency of the treatment effect under the discrete random effects model in multiregional clinical trials." *Stat Med*, 35:2301-14.
32. Negeri, Z. F., and Beyene, J. (2020), "Statistical methods for detecting outlying and influential studies in meta-analysis of diagnostic test accuracy studies." *Stat Methods Med Res*, 29:1227-1242.
33. Noma, H., Gosho, M., Ishii, R., Oba, K., and Furukawa, T. A. (2020), "Outlier detection and influence diagnostics in network meta-analysis." *Res Synth Methods*, 11:891-902.
34. Noma, H., Nagashima, K., Maruo, K., Gosho, M., and Furukawa, T. A. (2018), "Bartlett-type corrections and bootstrap adjustments of likelihood-based inference methods for network meta-analysis." *Stat Med*, 37:1178-1190.
35. Paule, R. C., and Mandel, J. (1989), "Consensus Values, Regressions, and Weighting Factors." *J Res Natl Inst Stand Technol*, 94:197-203.
36. Quan, H., Li, M., Shih, W. J., Ouyang, S. P., Chen, J., Zhang, J., and Zhao, P. L. (2013),

"Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials." *Stat Med*, 32:1691-706.

37. Quan, H., Mao, X., Chen, J., Shih, W. J., Ouyang, S. P., Zhang, J., Zhao, P. L., and Binkowitz, B. (2014), "Multi-regional clinical trial design and consistency assessment of treatment effects." *Stat Med*, 33:2191-205.
38. Quan, H., Mao, X., Tanaka, Y., Binkowitz, B., Li, G., Chen, J., Zhang, J., Zhao, P. L., Ouyang, S. P., and Chang, M. (2017), "Example-based illustrations of design, conduct, analysis and result interpretation of multi-regional clinical trials." *Contemp Clin Trials*, 58:13-22.
39. Quan, H., Zhao, P. L., Zhang, J., Roessner, M., and Aizawa, K. (2010a), "Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance." *Pharm Stat*, 9:100-12.
40. Quan, Hui, Li, Mingyu, Chen, Joshua, Gallo, Paul, Binkowitz, Bruce, Ibia, Ekapimo, Tanaka, Yoko, Ouyang, Soo Peter, Luo, Xiaolong, Li, Gang, Menjoge, Shailendra, Talerico, Steven, and Ikeda, Kimitoshi. (2010b), "Assessment of Consistency of Treatment Effects in Multiregional Clinical Trials." *Drug Information Journal*, 44:617-632.
41. Schmidt, Frank L., and Hunter, John E. 2015. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*.
42. Shimatani, Katsuyoshi, and Sudo, Takao. (2005), "<島谷(2005).pdf>." *Iryo To Shakai*, 15:43-51.
43. Sidik, K., and Jonkman, J. N. (2007), "A comparison of heterogeneity variance estimators in combining results of studies." *Stat Med*, 26:1964-81.
44. Sidik, Kurex, and Jonkman, Jeffrey N. (2005), "Simple heterogeneity variance estimation for meta-analysis." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54:367-384.
45. Stein, Markus Chagas, da Silva, Michel Ferreira, and Duczmal, Luiz Henrique. (2014), "Alternatives to the usual likelihood ratio test in mixed linear models." *Computational Statistics & Data Analysis*, 69:184-197.
46. Steyerberg, Ewout W., Harrell, Frank E., Borsboom, Gerard J. J. M., Eijkemans, M. J. C., Vergouwe, Yvonne, and Habbema, J. Dik F. (2001), "Internal validation of predictive models." *Journal of Clinical Epidemiology*, 54:774-781.

47. Teng, Zhaoyang, Lin, Jianchang, and Zhang, Bin. (2018), "Practical Recommendations for Regional Consistency Evaluation in Multi-Regional Clinical Trials with Different Endpoints." *Statistics in Biopharmaceutical Research*, 10:50-56.
48. Tohkin, M. (2016), "Regulatory science plays an important role in the global development of new drugs." *Nihon Yakurigaku Zasshi*, 148:18-21.
49. Tsou, H. H., James Hung, H. M., Chen, Y. M., Huang, W. S., Chang, W. J., and Hsiao, C. F. (2012), "Establishing consistency across all regions in a multi-regional clinical trial." *Pharm Stat*, 11:295-9.
50. Turpie, A. G., Lassen, M. R., Eriksson, B. I., Gent, M., Berkowitz, S. D., Misselwitz, F., Bandel, T. J., Homering, M., Westermeier, T., and Kakkar, A. K. (2011), "Rivaroxaban for the prevention of venous thromboembolism after hip or knee arthroplasty. Pooled analysis of four studies." *Thromb Haemost*, 105:444-53.
51. Uesaka, H. (2009), "Sample size allocation to regions in a multiregional trial." *J Biopharm Stat*, 19:580-94.
52. Ukyo, Yoshifumi, Noma, Hisashi, Maruo, Kazushi, and Gosho, Masahiko. (2019), "Improved Small Sample Inference Methods for a Mixed-Effects Model for Repeated Measures Approach in Incomplete Longitudinal Data Analysis." *Stats*, 2:174-188.
53. Veroniki, A. A., Jackson, D., Bender, R., Kuss, O., Langan, D., Higgins, J. P. T., Knapp, G., and Salanti, G. (2019), "Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis." *Res Synth Methods*, 10:23-43.
54. Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2016), "Methods to estimate the between-study variance and its uncertainty in meta-analysis." *Res Synth Methods*, 7:55-79.
55. Viechtbauer, W., and Cheung, M. W. (2010), "Outlier and influence diagnostics for meta-analysis." *Res Synth Methods*, 1:112-25.
56. Viechtbauer, Wolfgang. (2005), "Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model." *Journal of Educational and Behavioral Statistics*, 30:261-293.
57. Wedel, H., Demets, D., Deedwania, P., Fagerberg, B., Goldstein, S., Gottlieb, S., Hjalmarson, A., Kjekshus, J., Waagstein, F., Wikstrand, J., and Group, Merit-Hf Study. (2001), "Challenges of subgroup analyses in multinational clinical trials: experiences from

the MERIT-HF trial." *Am Heart J*, 142:502-11.
58. 厚生労働省. (2007), "国際共同治験に関する基本的考え方について."
59. 小山 暢之, 山本 英晴. (2016), "国際共同治験で留意すべき民族的要因について:—非弁膜性心房細動, 2 型糖尿病, 慢性閉塞性肺疾患, 双極性障害, 胃癌, 大腸癌を対象とした調査結果からの提言—." *レギュラトリーサイエンス学会誌*, 6:127-137.
60. 独立行政法人医薬品医療機器総合機構. (2006), "ロサルタン審査報告書."
61. 日本製薬工業協会. (2018), "国際共同治験での民族的要因の検討方法 3-layer approach の実践."
62. 汪 金芳, 田栗 正章. (1996), "ブ-トストラップ法 - 2 標本問題からの考察." *統計数理*, 44:3-18.

7. 付録

7.1 変量効果モデルにおける地域間分散パラメータの推定方法

7.1.1. Hedges-Olkin 推定量

Hedges と Olkin によって提案されたこの方法はコクラン推定量，又は分散成分型推定量とも呼ばれており，以下の式で表される。

$$\hat{\tau}^2 = \max\left\{0, \frac{1}{k-1}\sum_{i=1}^{k}(y_i - \bar{y})^2 - \frac{1}{k}\sum_{i=1}^{k}\sigma_i^2\right\} \tag{26}$$

ここで，$\bar{y}$はy_iの重み付けしない平均とする。Hedges-Olkin 推定量と DerSimonian-Laird 推定量の大きな違いは Hedges-Olkin 推定量では重み付けしない治療効果の分散に基づいているが，DerSimonian-Laird 推定量は重み付けした治療効果の分散に基づいている事である。この推定量は単純で反復計算も必要ない。DerSimonian-Laird 推定量，ML 推定量，REML 推定量と比較した時，τ^2が大きく，研究の数が多くなった時（$k \geq 30$），良い推定値となるが，大きい平均二乗誤差を持つ（Veroniki et al. 2016）。

7.1.2. Paule-Mandel 推定量

Q 統計量が自由度$k-1$のχ^2分布に従う事から，以下の式を$\hat{\tau}^2$に対して解く事により$\hat{\tau}^2$を推定する事が Paule と Mandel によって提案された。

$$Q(\hat{\tau}^2) = k - 1 \tag{27}$$

なお，$Q(0) < k-1$の場合には，$\hat{\tau}^2 = 0$となる。また，$\hat{\tau}^2$は反復計算により推定される。この推定量ではk及びτ^2が小さい場合に過大評価してしまい，大きい場合には過小評価となる。しかしながら，他の推定量に比べて，バイアスの程度が小さい事が知られている（Veroniki et al. 2016）。

7.1.3. Hunter-Schmidt 推定量

Hunter と Schmidt により提案された推定量は以下の式により与えられる。

$$\hat{\tau}^2 = \max\left\{0, \frac{Q-k}{\sum_{i=1}^{k} w_i}\right\} \tag{28}$$

この推定量では反復計算が必要ない。負のバイアスを持ち，その他の方法と同等又は低い平均二乗誤差を持つ事が知られている（Viechtbauer 2005）。なお，kが小さい時には特に過小評価してしまう事も知られている。

7.1.4. Sidik-Jonkman 推定量

Sidik と Jonkman により提案された推定量は以下の式により与えられる。

$$\hat{\tau}^2 = \frac{1}{k-1}\sum_{i=1}^{k} \hat{q}_i^{-1}\left(y_i - \hat{\mu}_{\hat{q}}\right)^2 \tag{29}$$

ここで，$\hat{q}_i = \hat{r}_i + 1$，$\hat{r}_i = \sigma_i^2/\hat{\tau}_0^2$，$\hat{\tau}_0^2 = \sum_{i=1}^{k}(y_i - \bar{y})^2/k$，$\hat{\mu}_{\hat{q}} = \sum_{i=1}^{k} \hat{q}_i^{-1} y_i / \sum_{i=1}^{k} \hat{q}_i^{-1}$とする。この推定量では常に正の値をとり，反復計算が必要ない。また，kとτ^2が大きい場合，DerSimonian-Laird 推定量より平均二乗誤差は小さく，バイアスも小さい事が知られている（Veroniki et al. 2016）。逆に，kとτ^2が小さい場合には，平均二乗誤差もバイアスも大きくなる。

7.1.5. 経験ベイズ推定量

ML 推定と類似したτ^2の推定法として，経験ベイズ推定量が提案されている。この推定法では以下の式の反復計算によりτ^2が推定される。

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} \hat{w}_i \left\{ (k/(k-1))\left(y_i - \hat{\theta}_{\hat{w}}\right)^2 - \sigma_i^2 \right\}}{\sum_{i=1}^{k} \hat{w}_i} \tag{30}$$

ここで，$\hat{\theta}_{\hat{w}} = \sum_{i=1}^{k} \hat{w}_i y_i / \sum_{i=1}^{k} \hat{w}_i$，$\hat{w}_i = 1/(\sigma_i^2 + \hat{\tau}^2)$とする。なお，ML 推定や REML 推定と同様にτ^2に初期値を与える必要があり，負の推定値とならないよう，与えられる初期値$\hat{\tau}$は正である必要がある。また，τ^2が小さい場合にはバイアスが小さい推定量となる（Sidik and Jonkman 2007）。

7.2 MERIT-HF 試験におけるその他の評価項目の評価

3.5.3 節に記載した通り，MERIT-HF 試験においては主要評価項目として，①全死亡，②全死亡と理由を問わない入院，の 2 つが設定されていた。また，副次評価項目として，全死亡と心不全による入院が設定されていた。3.5.3 節では主要評価項目①のみの結果を提示したが，参考までにその他の評価項目に対してもフォレストプロット及び提案法の結果を示す。

7.2.1 別の主要評価項目における結果

本節では 3.5.3 節で記載した主要評価項目とは別の主要評価項目である全死亡と理由を問わない入院をイベントとした結果を示す。まず，この評価項目に対する国別のオッズ比のフォレストプロットを図 7 に示した。なお，この主要評価項目の評価においては，3.5.3 節で記載した主要評価項目と異なり，フィンランドとスイスにおいてもイベントを発現しているため，解析に含めた。

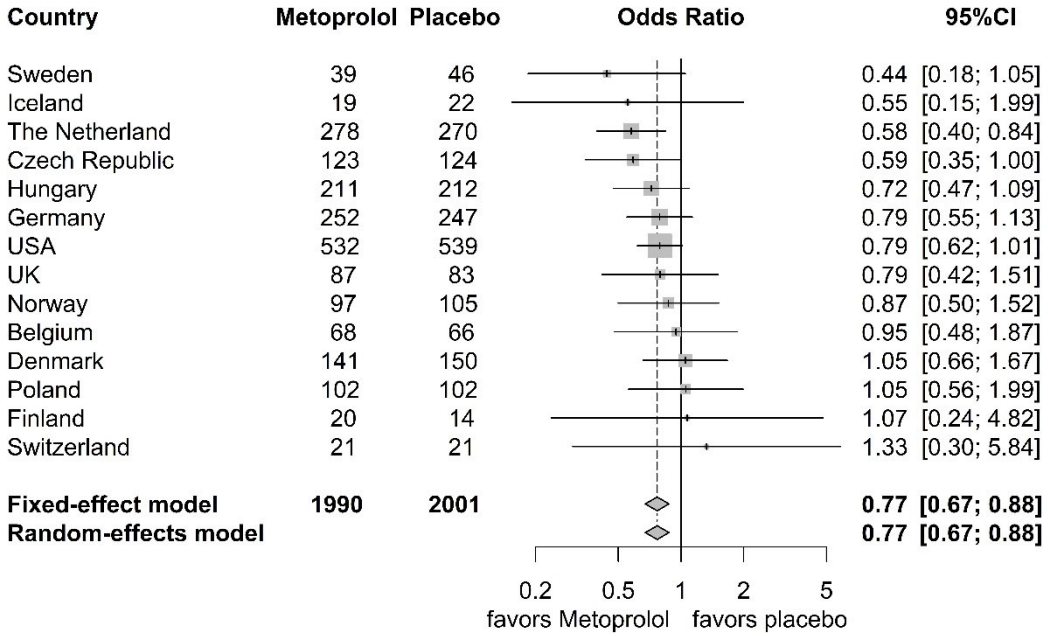

図 7 MERIT-HF 試験における全死亡と理由を問わない入院をイベントとした時の国別オッズ比のフォレストプロット

この評価項目における地域間の異質性の DerSimonian-Laird 推定量$\hat{\tau}^2$は 0 であり，提案法においては LOOCV 型のスチューデント化残差に基づく方法と尤度比検定に基づく方法のみを適用した。これより，LOOCV 型のスチューデント化残差に基づく方法と尤度比検定に基づく方法を適用した結果をそれぞれ表 23，表 24 に示す。いずれも外れ値として検出される国はなく，全死亡のみをイベントとした主要評価項目で外れ値として検出されたアメリカはここでは外れ値として検出されなかった。

表 23 MERIT-HF 試験における全死亡と理由を問わない入院をイベントした時の固定効果モデルの LOOCV 型のスチューデント化残差を用いた外れ値となる国の評価結果

国	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
オランダ	−1.584	−1.966	1.941
デンマーク	1.361	−1.969	1.990
スウェーデン	−1.265	−1.968	1.969
チェコ共和国	−1.009	−1.906	1.883

国	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
ポーランド	1.003	−1.928	1.916
スイス	0.729	−1.966	1.900
ベルギー	0.616	−1.952	1.872
アイスランド	−0.502	−1.937	2.036
ノルウェー	0.453	−2.066	1.844
フィンランド	0.437	−1.883	1.977
ハンガリー	−0.318	−2.051	2.011
アメリカ	0.280	−1.912	1.905
ドイツ	0.160	−1.910	2.002
イギリス	0.105	−2.000	1.906

表 24 MERIT-HF 試験における全死亡と理由を問わない入院をイベントした時の固定効果モデルの尤度比統計量を用いた外れ値となる国の評価結果

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
オランダ	2.510	4.049	0.116
デンマーク	1.852	3.813	0.174
スウェーデン	1.600	3.655	0.209
チェコ	1.019	4.075	0.317
ポーランド	1.005	3.914	0.323
スイス	0.531	3.861	0.484
ベルギー	0.379	3.726	0.540
アイスランド	0.252	3.892	0.621
ノルウェー	0.205	3.812	0.641
フィンランド	0.191	3.715	0.654

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
ハンガリー	0.101	4.104	0.752
アメリカ	0.078	4.029	0.779
ドイツ	0.026	3.685	0.864
イギリス	0.011	4.185	0.907

7.2.2　副次評価項目における結果

ここでは副次評価項目である全死亡と心不全による入院をイベントとした時の結果を示す。まず，この評価項目に対する各国のオッズ比のフォレストプロットを図 8 に示した。なお，この副次評価項目の評価においても，フィンランドとスイスはイベントを発現しているため，解析に含めた。

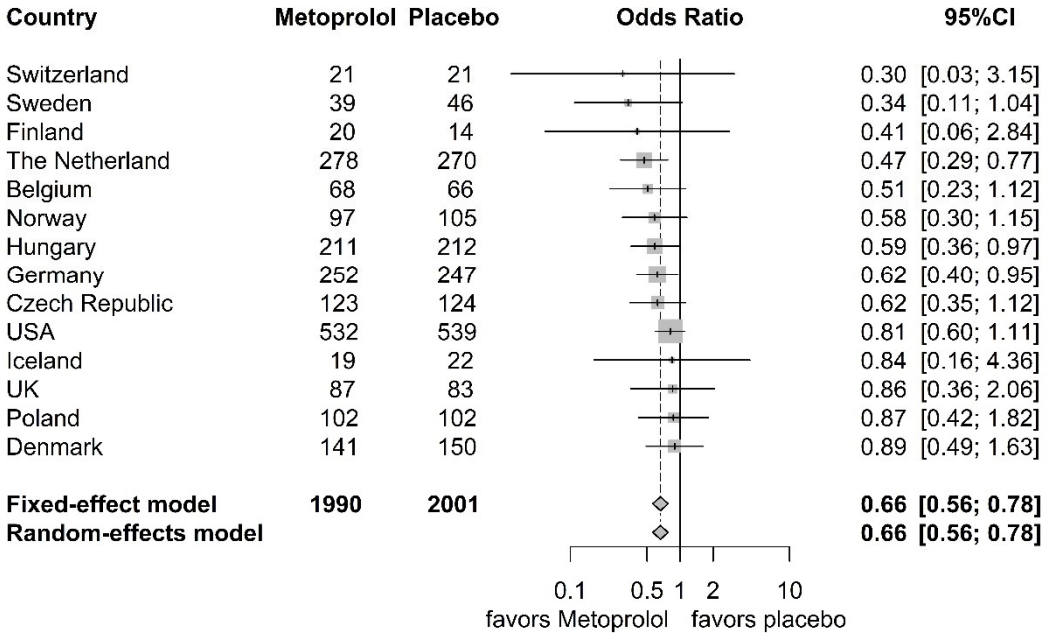

図 8 MERIT-HF 試験における全死亡と心不全による入院をイベントとした時の国別オッズ比のフォレストプロット

この評価項目における地域間の異質性の DerSimonian-Laird 推定量$\hat{\tau}^2$も 0 であり，提案法においては LOOCV 型のスチューデント化残差に基づく方法と尤度比検定に基づ

く方法のみを適用した。これより，LOOCV 型のスチューデント化残差に基づく方法と尤度比検定に基づく方法を適用した結果をそれぞれ表 25，表 26 に示す。いずれも外れ値として検出される国はなく，全死亡のみをイベントとした主要評価項目で外れ値として検出されたアメリカはここでも外れ値として検出されなかった。

表 25 MERIT-HF 試験における全死亡と心不全による入院をイベントした時の固定効果モデルの LOOCV 型のスチューデント化残差を用いた外れ値となる国の評価結果

国	t_i	ブートストラップ 2.5%点	ブートストラップ 97.5%点
アメリカ	1.543	−1.951	1.868
オランダ	−1.443	−1.911	2.008
スウェーデン	−1.189	−1.997	1.974
デンマーク	1.016	−1.911	1.891
ポーランド	0.738	−1.976	1.882
ベルギー	−0.674	−1.972	2.002
スイス	−0.662	−1.927	1.945
イギリス	0.583	−2.018	1.947
フィンランド	−0.492	−1.980	1.912
ハンガリー	−0.490	−1.936	2.075
ノルウェー	−0.371	−1.922	1.920
ドイツ	−0.338	−1.931	1.998
アイスランド	0.291	−2.011	1.955
チェコ	−0.209	−1.967	1.881

表 26 MERIT-HF 試験における全死亡と心不全による入院をイベントした時の固定効果モデルの尤度比統計量を用いた外れ値となる国の評価結果

国	尤度比統計量	ブートストラップ 95%点	ブートストラップ P 値
アメリカ	2.381	4.098	0.132
オランダ	2.083	3.794	0.144
スウェーデン	1.413	3.771	0.221
デンマーク	1.031	3.690	0.302
ポーランド	0.545	3.841	0.459
スイス	0.438	3.679	0.507
ベルギー	0.454	3.868	0.508
イギリス	0.340	4.097	0.564
ハンガリー	0.240	3.611	0.612
フィンランド	0.242	3.808	0.624
ノルウェー	0.138	3.552	0.714
ドイツ	0.114	4.018	0.727
アイスランド	0.085	4.177	0.775
チェコ	0.044	3.665	0.839