2020 ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop, Washington, DC, September, 2020

Methods for detecting outlying regions and influence diagnosis in multi-regional clinical trials

Makoto Aoki*
Department of Statistical Science, School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies, Tokyo, Japan
Integrated Biostatistics Department, Novartis Pharma K. K., Tokyo, Japan
*e-mail: m-aoki@ism.ac.jp

Hisashi Noma**
Department of Data Science, The Institute of Statistical Mathematics, Tokyo, Japan
**e-mail: noma@ism.ac.jp

Background and Objectives
Due to the globalization of drug development, multi-regional clinical trials (MRCTs) have been increasingly adopted in clinical evaluations. In MRCTs, the primary objective is to demonstrate the efficacy of new drugs in all participating regions [1,2]. However, because of the heterogeneity of various relevant factors across these regions, the treatment effects in different areas may not be consistent. Thus, assessing the consistency of treatment effects has become a relevant statistical issue in MRCTs [1,3,4]. In theoretical and applied statistics, effective methods for detecting extreme profiles and for determining how these profiles influence diagnostic methods have been well investigated [5], but they have not been well discussed in the context of MRCTs.
Results
In this article, we propose a set of novel influence diagnostic tools that can effectively identify outlying regions in MRCTs. The proposed influential diagnostic tools are (1) studentized residual obtained by a leave-one-out cross validation (LOOCV) scheme, (2) model-based approach using the likelihood ratio statistic, (3) relative change measure based on the variance of the overall treatment effect, and (4) relative change measure based on the between-region heterogeneity variance under the random-effects model. In addition, we propose to adapt bootstrap methods to assess the statistical significance and variabilities of these influence diagnostic tools. If outlying or influential regions are detected by the proposed methods and causes of regional difference can be identified, the treatment effect by the regional cluster can be quantified. We illustrate the effectiveness of the proposed methods by applying them to real MRCT. We show that some potential outlying regions were detected by the proposed methods. We also demonstrate that the overall results and their interpretations may be amended after exclusion of the detected influential regions.

Random effects model and fixed effects model for MRCTs

We discuss the influence diagnostic methods using region-specific summary statistics y_i $(i = 1, \cdots, k)$ obtained from subset analysis of the ith region. y_i corresponds to the treatment effect estimate of the ith region, e.g., the mean difference, log odds ratio and log hazard ratio. First, considering the common effect assumption across k regions, we assume the normal distribution model for y_i using the large sample approximation,

$$y_i \sim N(\theta, \sigma_i^2),$$

where θ is the common effect parameter and σ_i^2 is the region-specific variance obtained from the subset analyses. This model is well known in meta-analysis as the fixed-effect model [5]. Another standard statistical model is the random-effects model, which considers the between-regions heterogeneity [3,4],

$$y_i \sim N(\theta_i, \sigma_i^2),\ \theta_i \sim N(\mu, \tau^2), \qquad (*)$$

where μ is the grand mean for the treatment effects of k regions and τ^2 is the between-regions variance. We adopt the inverse-variance method for the fixed-effect model and the restricted maximum likelihood (REML) methods for the random-effects model as standard methods, but other estimation methods can also be used. We mainly explain the proposed methods based on the random-effects model (*), since the model (*) corresponds to the fixed-effect model when $\tau^2 = 0$.

Studentized residual

Regarding influence diagnostics, we first discuss a LOOCV-type influence measure that is similar to the dfbeta statistic in conventional regression diagnostics [6]. Conventionally, residual-based influence diagnostic measures have been defined as standardized by their standard errors to be comparable for all analysis units [6]. Let $\hat{\mu}^{(-i)}$, $\tau^{2(-i)}$ be the REML estimators from the random-effects model (*) based on a dataset of $k-1$ regions that excludes the ith region $(i = 1,2, \cdots, k)$. Then, the LOOCV studentized residual is defined as

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

where $Var[y_i - \hat{\mu}^{(-i)}] = (\hat{w}_i^{(-i)})^{-1} + (\sum_{j \neq i} \hat{w}_j^{(-i)})^{-1}$ and $\hat{w}_j^{(-i)} = (\hat{\tau}^{2(-i)} + \hat{\sigma}_j^2)^{-1}$ $(j = 1,2, \cdots, k)$. Note that $\hat{\mu}^{(-i)}$ and $\hat{\tau}^{2(-i)}$ are estimated by the dataset of $k-1$ regions that excludes the ith region. Thus, t_i is interpreted as a predicted studentized residual of the ith region from the estimated random-effects model (*) by the other $k-1$ regions. For assessing the influences, a reference threshold can be obtained by the sampling distribution of t_i. t_i follows the standard normal distribution if the assumed random-effects model (*) is correct. Thus, a widely-used criterion is comparing 1.96 with the absolute value of t_i. If the criterion is fulfilled, the corresponding region might be considered as a potential outlier that exceeds the range of random variation. However, the criterion depends on the large-sample assumption, and the sampling variation might not be adequately quantified. Thus, we propose an alternative effective approach by adopting the parametric bootstrapping method to evaluate the sampling variation. The bootstrap algorithm is given as follows:

Algorithm (Parametric bootstrap for estimating the sampling distribution of t_i).
1. For the random effects model (*), compute the REML estimates of $\{\mu, \tau^2\}$.
2. Resample $y_1^{(b)}, y_2^{(b)}, \cdots, y_k^{(b)}$ from the estimated distribution of (*) with the parameters substituted with $\{\mu, \tau^2\}$ via parametric bootstrap B times $(b = 1, 2, \cdots, B)$.
3. Compute the LOOCV studentized residuals $t_i^{(b)}$ $(i = 1,2, \cdots, k)$ for the bth bootstrap sample $y_1^{(b)}, y_2^{(b)}, \cdots, y_k^{(b)}$; replicate it for all B bootstrap samples.
4. Obtain the bootstrap estimate of the sampling distribution of t_i by the empirical distribution of $t_i^{(1)}, t_i^{(2)}, \cdots, t_i^{(B)}$.

The 2.5th and 97.5th percentiles in this empirical distribution can be adopted as the thresholds to detect the potential outliers. The detected regions would be considered to be influential outliers that exceed the range of random variation. Note that the above methods can be similarly applied to the fixed-effect model by fixing τ^2 to be 0 and interpreting μ as the common effect parameter.

Model-based approach using the likelihood ratio statistic

The second approach is a model-based likelihood ratio test using a mean-shifted model that was originally proposed for meta-analysis by Negeri and Beyene [7] and Noma et al. [8]. The mean-shifted model assumes that the treatment effect of a participating region is different from the overall effect. For the random-effects model (*), we assume that the random-effect distribution for the corresponding jth region is shifted as $\theta_j \sim N(\mu + \zeta, \tau^2)$, and that of the other $k-1$ regions is the same as (*), i.e., $\theta_i \sim N(\mu, \tau^2)$ $(i \neq j)$. Then, we consider the following testing problem:

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

When the null hypothesis is rejected, the treatment effect of the ith region is significantly diverged from the overall mean and is a potentially outlying region.
For this testing problem, we consider a likelihood ratio test. The log likelihood function under the null hypothesis corresponds to that of the random-effects model (*) and is written as

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

Further, the log likelihood function under the alternative hypothesis is

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

Then, the likelihood ratio statistic is given as

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

where $\{\tilde{\mu}, \tilde{\tau}^2\}$ is the maximum likelihood (ML) estimate of the null model (*) and $\{\hat{\mu}_{[j]}, \hat{\tau}^2_{[j]}, \hat{\zeta}_{[j]}\}$ is the ML estimate of the mean-shifted model for the jth region. The likelihood ratio statistic $T_{[j]}$ approximately follows the χ^2 distribution with 1 degree of freedom under the null hypothesis. Therefore, for the 5% significance level test, we can adopt 3.84, which is the 95th percentile of the χ^2 distribution with 1 degree of freedom. Note that the model-based approach can be applied to the fixed-effect model by fixing τ^2 to be 0 for the above models and interpreting μ as the common effect parameter. Similar to the discussions in the method by the studentized residual, the large-sample approximation might be violated under realistic situations, and the chi-square approximation would potentially be invalid [8,9]. For this model-based approach, the bootstrap method is also applicable.

Relative change of the variance of the overall estimator of μ

Another popular approach for influence diagnostics is assessing the relative change of the variance estimate of the overall estimator of μ in the LOOCV scheme. Viechtbauer and Cheung [10] proposed this approach for the random-effect model of meta-analysis. It can be directly applied to MRCT settings based on the random-effects model (*). The influential statistic for the jth region, given as

$$VRATIO_j = \frac{Var[\hat{\mu}^{(-j)}]}{Var[\hat{\mu}]} = \frac{\sum_{i=1}^{N} \hat{w}_i}{\sum_{i \neq j} \hat{w}_i},$$

assesses the relative change of variances between a leave-one-region-out dataset and the dataset containing all participating regions. $VRATIO_j$ reflects the impact of the jth region on the precision of the overall treatment effect estimator and values on $(0, \infty)$. When $VRATIO_j$ is smaller than 1, the inclusion of the jth region gains the variation of the overall estimator, although the sample size is increased; this inclusion also usually enlarges the between-studies heterogeneity. In this case, the jth region might be an outlier that deviates from the overall population. For determining the criteria used to define outliers, the parametric bootstrap approach can be applied to estimate the sampling distribution of $VRATIO_j$ by substituting t_i for $VRATIO_j$ in Algorithm showed for the standardized residual. For a 5% significance level, the lower 5th percentile of the bootstrap distribution is used for the critical value.
Note that if we adopt the fixed-effect model, the variance ratio statistic usually reflects the information sizes (i.e., simply proportional to sample sizes) of individual regions, and therefore it might not be an adequate influential measure to detect outlying regions. Thus, it should be applied to the analyses using the random-effects model (*).

Relative change of the heterogeneity variance τ^2

An influence measure similar to $VRATIO_j$ is considered for the heterogeneity variance estimates of the random-effects model (*). Viechtbauer and Cheung [10] proposed using the ratio of the estimates of τ^2 for the leave-one-trial-out dataset and the all-trial dataset, i.e.,

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

$TRATIO_j$ also values on $(0, \infty)$. The interpretation of $TRATIO_j$ is similar to that of $VRATIO_j$. If the $TRATIO_j$ is smaller than 1, the exclusion of the jth region decreases the heterogeneity among the population and can be seen as an outlying region that has an extreme profile relative to the overall population. For determining the criterion for outliers, the parametric bootstrap approach can be used to estimate the sampling distribution of $TRATIO_j$ by substituting t_i for $TRATIO_j$ in Algorithm showed for the standardized residual. For a 5% significance level, the lower 5th percentile of the bootstrap distribution is used for the critical value. Note that this measure cannot be defined for the fixed-effect model, and can be adopted only for the random-effects model analyses.

Application

We illustrate the proposed methods via applications to analysis of real MRCT. The dataset is from the RENAAL study, an MRCT of losartan in patients with type 2 diabetes and nephropathy [11,12]. Twenty-eight countries joined the RENAAL study, and the final analysis included 751 participants in the losartan group and 762 participants in the placebo group. The primary endpoint was a composite endpoint that consisted of a doubling of the baseline serum creatinine concentration, end-stage renal disease, or death. The results of a subgroup analysis of four participating regions (Asia, Latin America, Europe, and North America) are presented in Figure 1. There was substantial heterogeneity (τ^2=0.041). Asia accounts for about 15% of the overall participants, and it therefore might strongly influence the overall estimate of the treatment effect. We assessed its influence quantitatively by the proposed methods. First, we present the LOOCV studentized residuals in Table 1; the bootstrap percentiles were computed by 2400 resamples. As expected, the studentized residual t_i of Asia exceeded the bootstrap lower 2.5th percentile in both the fixed-effect model and the random-effects model. Second, in Table 2 we show the results of the model-based approach using the likelihood ratio statistic; the number of resamples was 2400. Then, in Table 3 we present the results of analyses of the relative change measures of the variance estimate of μ and the heterogeneity variance estimate; the number of resamples was 2400. As a whole, only the Asia region was consistently detected as a potential outlying region that would exceed the range of random variations obtained by all four of the proposed methods.
To evaluate sensitivity, we conducted a leave-one-out analysis of the Asia region. The hazard ratio (HR) estimate was 0.94 (95%CI: 0.79 to 1.11; p = 0.459) for the fixed-effect model and the heterogeneity variance estimate of the random-effects model became 0; both models provided identical results. The HR estimate was markedly changed and the significant difference was no longer significant. Thus, the Asia region may have had a strong influence on the overall effect estimate of this MRCT and could have influenced the primary conclusions. In contrast, the HR estimates in the other regions ranged from 0.91 to 0.95, which might indicate smaller effects compared with the overall results. The inconsistency of these estimates is intuitively clear, and our proposed methods explicitly detected the Asia region as influential by certain statistical criteria. As illustrated in this case, the proposed influence diagnostic methods should provide effective statistical evidence for assessing the influence of outlying regions in MRCTs.

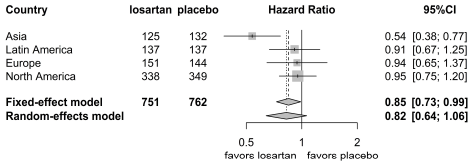

Figure 1. Forest plot of region-specific hazard ratio estimates in the RENAAL study

Table 1. Results of evaluation of outlying regions using the LOOCV studentized residual for the fixed-effect model and the random-effects model in the RENAAL study

Fixed-effect model				Random-effects model			
Region	t_i	Bootstrap lower 2.5th percentile	Bootstrap upper 2.5th percentile	Region	t_i	Bootstrap lower 2.5th percentile	Bootstrap upper 2.5th percentile
Asia	−2.796	−2.058	1.991	Asia	−2.792	−2.478	2.561
North America	1.290	−1.924	1.912	North America	0.612	−2.808	2.873
Europe	0.599	−2.032	2.002	Europe	0.475	−2.421	2.334
Latin America	0.532	−1.907	1.921	Latin America	0.389	−2.810	2.583

Table 2. Results of evaluation of outlying regions using the likelihood ratio (LR) statistics for the fixed-effect model and the random-effects model in the RENAAL study

Fixed-effect model				Random-effects model			
Region	LR statistic	Bootstrap 95th percentile	Bootstrap p-value	Region	LR statistic	Bootstrap 95th percentile	Bootstrap p-value
Asia	7.815	3.889	0.009	Asia	6.935	5.013	0.014
North America	1.663	3.725	0.182	North America	0.899	5.381	0.476
Europe	0.358	4.090	0.544	Europe	0.339	5.035	0.624
Latin America	0.283	3.676	0.608	Latin America	0.257	4.836	0.651

Table 3. Results of evaluation of outlying regions using relative change measures for the variance of $\hat{\mu}$ and the heterogeneity variance $\boldsymbol{\tau^2}$ in the RENAAL study

Relative change in variance of $\hat{\mu}$			Relative change in estimation of heterogeneity parameter $\boldsymbol{\tau^2}$		
Region	VRATIO	Bootstrap lower 5th percentile	Region	TRATIO	Bootstrap lower 5th percentile
Asia	0.418	0.461	Asia	0.000	0.003
Europe	1.788	0.448	North America	1.529	0.001
North America	1.877	0.521	Europe	1.608	0.002
Latin America	1.954	0.462	Latin America	1.733	0.002

References

1. Quan H, Li M, Shih WJ, et al. Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. Stat Med. 2013; 32(10):1691-1706.
2. International Conference Harmonization. General Principles for Planning and Design of Multi-Regional Clinical Trials. 2017; Available at: https://database.ich.org/sites/default/files/E17EWG_Step4_2017_1116.pdf.
3. Quan H, Zhao PL, Zhang J, et al. Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. Pharm Stat. 2010; 9(2):100-112.
4. Liu JT, Tsou HH, Gordon Lan KK, et al. Assessing the consistency of the treatment effect under the discrete random effects model in multiregional clinical trials. Stat Med. 2016; 35(14):2301-2314.
5. Higgins J, Thomas J. Cochrane Handbook for Systematic Reviews of Interventions. 2nd edn. Chichester: Wiley; 2019.
6. Belsley DA, Kuh E, Welsch RE. Regression Diagnostics : Identifying Influential Data and Sources of Collinearity. New York: Wiley; 1980.
7. Negeri ZF, Beyene J. Statistical methods for detecting outlying and influential studies in meta-analysis of diagnostic test accuracy studies. Stat Methods Med Res. 2019; DOI: 10.1177/0962280219852747.
8. Noma H, Gosho M, Ishii R, et al. Outlier detection and influence diagnostics in network meta-analysis. 2019; arXiv: 1910.13080.
9. Veroniki AA, Jackson D, Bender R, et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. Res Synth Methods. 2019; 10(1):23-43.
10. Viechtbauer W, Cheung M W. Outlier and influence diagnostics for meta-analysis. Res Synth Methods. 2010; 1(2):112-125.
11. Brenner BM, Cooper ME, de Zeeuw D, et al. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. N Engl J Med. 2001; 345(12):861-869.
12. Pharmaceuticals and Medical Devices Agency. Review report for Losartan. 2006; Available at: pmda.go.jp/drugs/2006/P200600021/63015300_21000AMZ00678_Q101_1.pdf.