International differences in treatment effect: do they really exist and why?†

Stuart Pocock1*, Gonzalo Calvo2, Jaume Marrugat3, Krishna Prasad4, Luigi Tavazzi5, Lars Wallentin6, Faiez Zannad7, and Angeles Alonso Garcia8

1Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK; 2Hospital Clinic of Barcelona, Barcelona, Spain; 3Research Group on Cardiovascular Epidemiology and Genetics, IMIM, Barcelona, Spain; 4Medicines and Healthcare Products Regulatory Agency and Guy’s and St Thomas’ Hospital, London, UK; 5Maria Cecilia Hospital, GVM Care & Research, Cotignola, Italy; 6Uppsala Clinical Research Centre, Uppsala University, Uppsala, Sweden; 7INSERM, Centre Hospitalier Universitaire, Nancy, France; and 8Hospital Universitario Puerta de Hierro, Madrid, Spain

Received 1 September 2012; revised 14 January 2013; accepted 8 February 2013; online publish-ahead-of-print 7 March 2013

With the increasing globalization of clinical trials, the opportunity exists to explore potential geographic differences in treatment effect within any major trial. Such geographic differences may arise because of international differences in patient selection, medical practice, or evaluation of outcomes, and such international variations need better documentation in trial reports. Appropriate pre-defined statistical analyses, including statistical tests of interaction regarding geographic heterogeneity in treatment effect, are important. Geographic variations are a particularly tricky form of subgroup analysis: they lack statistical power, are at best hypothesis-generating and can generate more confusion than insight. Referring to key examples, e.g. the PLATO and MERIT-HF, we emphasize the need for caution in interpreting evidence of potential geographic inconsistencies in treatment effect. Although it is appropriate to explore any biological or practical reasons for apparent geographic anomalies in treatment effect, the play of chance is often the most plausible and wise interpretation.

Keywords: Randomized clinical trials • Geographic differences • Subgroup analysis

Introduction

Randomized clinical trials of cardiovascular diseases are increasingly recruiting patients globally in order to achieve the size necessary to provide reliable, conclusive overall findings on treatment efficacy and safety. Such global recruitment also enhances the generalizability of a trial’s findings to a broad international population of patients.

In principle, such international recruitment provides the opportunity to explore potential geographic variations in treatment effect. The problem is that any trial is not adequately powered to produce reliable evidence from such geographic subgroup analyses. Hence, for most trials, one does not detect any meaningful geographic inconsistencies and the overall result is deemed to be universally applicable.

However, occasionally, a trial does produce an apparent geographic anomaly, which then produces intense debate about whether the geographic variations in treatment effect truly exist beyond the play of chance, and if so what is the explanation.1,2 The PLATO trial of ticagrelor in ACS is the latest puzzling finding,3,4 whereas the MERIT trial of metoprolol in heart failure provoked a similar uproar 10 years earlier.5,6 In both instances, the overall superiority of the new treatment was not evident in US patients, which had important implications for whether the FDA should approve the drug’s use in the USA.

This article’s aim is to provide a cautious perspective on how we should explore and interpret potential geographic variations in treatment effect, taking into account the statistical limitations of subgroup analyses, the possibility (or not) of genuine geographic disparities existing, and the need to exercise collective wise judgement when unexpected findings arise.

Why should geographic variations exist?

It is highly unlikely that for any major international trial, all countries are identical as regards the type of patients recruited, their therapeutic management, and the evaluation of their outcomes. The question is whether such international discrepancies can be

---

†This article was stimulated by an ESC Regulatory Workshop on the topic held on 10 February 2012 by the European Society of Cardiology. The CRT (Cardiovascular Round Table) is a strategic forum for high-level dialogues between industry and ESC leadership to identify and discuss key strategic issues for the future of cardiovascular health in Europe.

* Corresponding author. Tel. +44 2079272413, Fax +44 2076372853, Email: stuart.pocock@lshtm.ac.uk

Published on behalf of the European Society of Cardiology. All rights reserved. © The Author 2013. For permissions please email: journals.permissions@oup.com
of sufficient magnitude to seriously generate real geographic differences in treatment efficacy and safety that are clinically important. First, let us consider what types of geographic variation might exist.

**Patient selection**

There may exist innate ethnic (e.g. genetic) or environmental differences which result in varying disease incidence and severity that could in principle influence how a treatment affects its prognosis in different regions. Such innate differences could also affect the risk/benefit ratio, e.g. if certain ethnic groups have a different side-effect profile. While ethnic or genetic differences are rarely responsible for geographic differences in treatment effect they can arise, e.g. (i) the dose of warfarin, prasugel, and beta-blockers is different for Asian subjects and (ii) the differences in incidence and mortality of coronary heart diseases by country observed in Europe, which may imply that a distinct benefit/risk balance might exist. Also, differences in the cost-effectiveness of an intervention can stem from the fact that relative risks for an intervention may be similar in different countries/regions but the absolute effect or the absolute risk may be quite different. This is the case for total cholesterol or systolic blood pressure illustrated by the French paradox and the results of the Seven Countries Study published a few years ago.

Differences in health-care systems and other practicalities of how patients present and how clinical teams select whom to include in a trial may generate real between-country differences in patient characteristics. Specifically, if a country (or region) recruits higher risk patients, then, even if the relative benefits of a new treatment are geographically consistent, the absolute benefit would be greater in such a country.

**Medical practices**

There will undoubtedly be international differences in patient management such that ancillary patient care including other (non-randomized) drug and interventional treatments may be markedly differing across countries. The EVEREST study is an example where the use of established pharmacotherapy for heart failure (aldosterone antagonists) differed considerably.

Furthermore, the actual delivery of the randomized treatments may differ between countries, despite attempts in the study protocol to be specific on what is intended. For instance, the approach to dose modifications, up-titrations (if present), and rescue medications may vary between centres, as may patient compliance with drug regimens and the incidence of patient withdrawals from randomized treatment. One recent instance concerns trials of new anticoagulants in atrial fibrillation which revealed marked geographic variations in the quality of warfarin treatment in the control arm. Another instance concerns a variable use of interventional treatments, e.g. coronary revascularization in trials of medical treatments of coronary artery disease. As coronary revascularization substantially affects the event rates, the efficacy of pharmacological interventions will vary in relation to the utilization and timing of such procedures.

All these international inequalities in treatment practices could affect how the true efficacy and safety of a new treatment manifests itself in different countries.

**Evaluation of outcomes**

Most trial protocols go to considerable lengths to ensure that patient outcomes (endpoints) are very precisely defined using objective criteria. Nevertheless, investigators in different countries may interpret these criteria differently. The existence of an independent, blinded Event Validation Committee will enhance geographical consistency, but still were any centres to under-report potential events this could still be a problem. For instance, MI events could be missed if appropriate biomarkers are not drawn routinely in certain countries.

Also, certain events, e.g. hospitalization for a specific outcome and revascularization procedures (planned or unplanned), could have different thresholds for their occurrence, depending on the health-care system and local practices in different countries. In particular, trials in chronic heart failure have heart failure hospitalization as a key component of a composite endpoint, including mortality, though the inherent fragility/inconsistency of this endpoint across countries is well known. In general, harder endpoints (e.g. cardiovascular death or admission for an ACS) are easier to collect consistently, whereas softer endpoints are more prone to differences in recognition and reporting.

A further issue is that the quality of trial conduct may vary across countries, with some contributing more protocol violations, missing data, and loss to follow-up than others. A different problem is when the findings from a trial performed wholly in one region are extrapolated to clinical practice in other regions. For instance, the large COMMIT trial conducted wholly in China formed the basis for claim for extension of the indication for clopidogrel in ACS in Europe.

The FDA often requests that a pivotal trial recruit a certain minimum number of patients in the USA (typically 20–30%), but this is sometimes difficult to achieve. Also, one might question how helpful it is in practice since such a US subgroup may still be too small to provide clear evidence specific to the US population.

The European Medicines Agency recognizes that an increasing percentage of pivotal studies submitted to EU regulatory authorities are conducted outside the EU and there is a need to understand the differences and concerns that may arise in the extrapolation of study results to the EU population. Experience so far has shown that intrinsic as well as extrinsic factors are important to consider when extrapolating data. With an expanding EU, differences in terms of extrinsic and intrinsic factors also increase within the EU. In particular, extrinsic factors, such as medical practice, disease definition, and study population, may influence the applicability of such data sets. Prospective analysis of potential extrinsic and/or intrinsic factors when conducting a clinical trial in a certain region may help regulatory assessors to evaluate whether certain clinical trials conducted in a specific area of the world are relevant to the EU setting or if there are reasons to perform additional clinical trials within the EU.

**Documenting geographic differences in practice**

Before we move on to detecting potential geographic differences in treatment effects, we feel trialists should put more effort into
documenting geographic differences in patient characteristics, patient management and patient outcomes for both randomized treatment groups combined. Such data by geographic region might be too extensive to routinely include in published trial reports, but should be available for the trialists themselves to explore geographic variations in their study sample of patients, or in supplementary tables available online.

For instance, it would be helpful to examine the following by geographic region (for both treatment groups combined):

(i) key baseline characteristics of patients, especially for variables that are known to be strongly associated with prognosis;
(ii) concomitant medications (e.g. use of inotropic agents in heart failure) and interventional procedures (use of angiography and revascularization for ACS) are also relevant in affecting prognosis and the impact of a new treatment;
(iii) measures of patient compliance, and the extent of both patient withdrawal from randomized treatment and loss to follow-up;
(iv) primary and key secondary endpoints, and their individual components if composites are used.

The latter is valuable in assessing whether prognosis varies markedly by geographic region. If this is the case (or anticipated as plausible), then analyses that adjust for region (and other key baseline predictors) should be undertaken as a secondary exploration.

Region can be a somewhat artificial grouping of countries (e.g. the heterogeneity of cultures and health systems in Europe, though the expanding EU may have some harmonizing effect), so that if data are sufficiently large such exploration of geographic patterns could also be done by country (or even by centre among major recruiters).

Assessing the strength of evidence for geographic variations in treatment effect

Geographic region is a subgroup, and hence the usual principles of how to perform and interpret subgroup analyses apply to this situation. Pre-defined subgroups are preferred, and hence the first desirable step is to precisely pre-declare the exact groupings of countries into regions. This is not always easy, since one juggles geographic cohesiveness with the need to have enough patients in each region to merit their separate analysis. For instance, for European patients, does one form separate regions for eastern and western European countries, or can one subdivide yet further into northern and southern countries of western Europe? No grouping of countries is completely satisfactory, but sensible attempts prior to analysis are better than post hoc groupings that carry the risk of artificially inflating apparent geographic

Figure 1  Estimated treatment effects by geographic region for the primary endpoint (CV death, MI, or stroke) of the PLATO trial (hazard ratios with 95% CIs, interaction P-value 0.05).
disparities. Consideration should be given to countries’ similarity in practice patterns rather than just being geographic neighbours.

The next step is to display the primary outcome’s treatment difference by geographic region in graphical or tabular form. This is best done as a forest plot, displaying the odds ratio or hazard ratio (and its 95% confidence interval) comparing new treatment with standard (placebo or active control). Alongside, it is helpful to tabulate by both treatment group and region the numbers of patients and also the numbers (or percentage) experiencing the primary event outcome. In addition, all these items should be presented for the whole study, i.e. all regions combined.

Visual inspection of these data can be revealing. For instance, if all confidence intervals are overlapping, then it is wise to conclude there is little evidence of geographical variations in treatment effect. On the other hand, if one region’s estimate is distinctly different from the others, then it appears a geographical disparity may be present.

Figure 1 presents such findings for the PLATO trial’s primary endpoint (cardiovascular death, myocardial infarction of stroke) comparing ticagrelor with clopidogrel in 18 624 patients with acute coronary syndrome.3,4 The result is provocative since although the overall treatment difference is strongly in favour of ticagrelor, for patients in North America there is a weak tendency in the opposite direction.

So how strong is the evidence for geographic heterogeneity here? The most useful objective guide is a statistical test of interaction which directly assesses whether the variation in hazard ratios across the four regions could plausibly have arisen by chance. In this case, interaction $P = 0.05$, which means there is some evidence of geographic disparity, but such borderline significance leaves the matter open to doubt.

If one judges that genuine geographic heterogeneity may well exist, then an alternative random-effects model may be used to capture such regional differences. By its incorporation of potential treatment by region interaction effects, the random-effects model will tend to widen the confidence interval for the overall treatment-effect estimate. This appropriately expresses the increased uncertainty of any overall estimate in the presence of geographical heterogeneity. However, problems are deciding on which level of detail (region, country, or site) to represent with this random-effect, and also any small outlier geographic component tends to receive more weight in the analysis and may distort findings compared with analyses stratified by region or country.

### Interpreting apparent geographic variations and exploring possible reasons

As with other subgroups analyses, great caution is needed in interpreting signs of geographic heterogeneity in treatment effect. Clinical trials are powered to give clear evidence of an overall treatment effect, and hence tend to lack power for reliable exploration of regional (and other subgroup) differences.

Hence, subgroup analyses are generally seen as exploratory and hypothesis-generating, so that unless the evidence of heterogeneity is overwhelming (e.g. interaction $P < 0.001$) they cannot lead to a definitive conclusion.

In the PLATO trial, the between-region comparison was one of 32 pre-planned subgroup analyses, and hence purely by chance one could expect one or two such analyses to have interaction $P < 0.05$. This should act as a restraint on any assertive claims of genuine subgroup effects. Furthermore, post hoc emphasis on the most striking subgroup finding (geography, in this case) means that even if the finding is not entirely due to chance, the observed data are prone to exaggerate any true disparities (between regions).

It is also helpful to place any trial’s evidence in the broader context of international trial’s research in general. For instance, each year there are several major cardiovascular international randomized trials reported, but convincing evidence of geographic variation is rarely spotted. For instance, the last trial we recall that provoked such hot debate on geographic differences was the MERIT-HF trial of metoprolol in heart failure.5,6 First published in 1999. Although the overall trial evidence showed a highly significant mortality reduction on metoprolol compared with placebo, for the subgroup of US patients no mortality difference existed (see Table 1). After much detailed data exploration, no sensible explanation was found. This led one enquiry5 to conclude: ‘We should expect some variation of the treatment-effect around the overall estimate as we examine a large number of subgroups because of small sample size in subgroups and chance. Thus the

### Table 1 International variations in mortality by treatment group in the MERIT-HF trial

<table>
<thead>
<tr>
<th>Metoprolol</th>
<th>Placebo</th>
<th>Hazard ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA vs. the rest (post hoc interaction test $P = 0.003$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1990</td>
<td>2001</td>
</tr>
<tr>
<td>Deaths</td>
<td>145</td>
<td>217</td>
</tr>
<tr>
<td>USA</td>
<td>51</td>
<td>49</td>
</tr>
<tr>
<td>Other countries</td>
<td>94</td>
<td>168</td>
</tr>
<tr>
<td>Deaths by country (country by treatment interaction test $P = 0.22$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungary</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>Germany</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td>Netherlands</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>Belgium</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>Sweden</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Norway</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>UK</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Finland</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Iceland</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Poland</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Denmark</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>USA</td>
<td>51</td>
<td>49</td>
</tr>
</tbody>
</table>
best estimate of the treatment effect on total mortality for any subgroup is the estimate of the hazard ratio for the overall trial.

A subsequent meta-analysis of heart failure beta-blocker trials has claimed that beta-blockade has a lower magnitude of survival benefit in the USA than the rest of the world, but an editorial comment concluded that the analyses are provocative and possibly hypothesis generating, but they should not be interpreted as demonstrating a lesser beta-blocker benefit in North American patients.

So, across the spectrum of all major international cardiovascular trials, we appear to generate one tantalizing geographic finding less than once every 10 years.

Thus, one’s prior belief that any specific trial should truly exhibit geographic discrepancies in treatment effect is inevitably low, and furthermore one has no clear prior view as to the direction (or specific affected regions) of such variations.

The difficulties in reliably identifying any true geographic differences in treatment effect are illustrated in the appendix. This simulation exercise documents how attempts at exploring such geographic heterogeneity are seriously impaired by lack of statistical power.

**Post hoc analyses: playing with P-values**

Geography is a particular challenging form of subgroup analysis since there are many different ways of forming geographic groupings, each yielding a rather different strength of evidence for geographic heterogeneity.

For instance, in the PLATO trial, if, instead of comparing four regions (as in Figure 1), one compares the USA with all other countries combined, their hazard ratios are, respectively, 1.27 (95% CI 0.92–1.75) and 0.81 (95% CI 0.74–0.90) with interaction $P = 0.009$. Alternatively, one can assess all 43 countries separately, and the global interaction test for heterogeneity among the 43 hazard ratios yields $P = 0.95$. Furthermore, the observed hazard ratio exceeded 1 in 12 countries, but given only chance variation (i.e. no real heterogeneity), one would have expected this to occur in 13 countries anyway.

Such post hoc analyses of geographic variations can be twisted to present either a more positive or a more negative impression of heterogeneity, depending on the option chosen. Thus, it is wise to concentrate on the original pre-defined regional differences (Figure 1) as the unbiased evidence of potential heterogeneity.

**Exploring why true geographic differences may exist**

If the evidence for a true geographic variation in the treatment effect is deemed sufficiently convincing, then one needs to explore possible explanations. The section ‘Why should geographic variations exist’ referred to three general considerations: differences in patient selection, treatment practice, and evaluation of outcomes. The difficulty here is the lack of any specific prior hypothesis, so one is inevitably drawn into exploring a multiplicity of potential explanatory factors.

For instance, in the PLATO trial, the investigators declare that 37 such candidate explanatory variables were investigated: 20 baseline characteristics, 8 ancillary treatments at baseline, and 9 aspects of care/medication post-randomization. In addition, various aspects of study quality/performance were qualitatively assessed. The emphasis was on comparing US patients with the rest of the world. None statistically explained the regional interaction except for the maintenance dose of aspirin.

Fifty-four per cent of US patients received a median maintenance aspirin dose $\geq 300$ mg compared with 1.7% of non-US patients. Furthermore, for post hoc subgroups of high ($\geq 300$ mg) and low ($<300$ mg) aspirin maintenance dose, the primary end-point hazard ratios for ticagrelor vs. clopidogrel were 0.79 (95% CI 0.71–0.88) and 1.45 (95% CI 1.01–2.09), respectively, with interaction $P = 0.0006$. This interaction of treatment with aspirin (shown in Figure 2) is sufficiently marked to statistically explain the apparent US discrepancy in treatment effect.

At face value, this looks a convincing explanation of the US anomaly and has led the FDA to issue a boxed warning ‘use of ticagrelor with aspirin doses exceeding 100 mg/day decreases its effectiveness’. But one needs to remember that geographic region was one of 32 pre-defined subgroup analyses, which was then re-grouped into US vs. non-US patients and for which at least 37 possible explanatory variables were explored. A sceptic could argue that in the complete absence of any real subgroup phenomena, a post hoc result as seemingly impressive as this could plausibly be generated by the play of chance alone.

**Conclusions**

It is plausible to argue that any overall benefit of a new treatment may not apply equally to all eligible patients, and the concept of ‘personalized medicine’ is intuitively attractive. Regulators have a responsibility to interpret evidence from subgroups, not necessarily to restrict indications but to protect subpopulations from unacceptable risk (e.g. contraindication of prasugrel in patients with a history of stroke). Geographic variations are one component of this general principle. Like all other subgroup analyses, they are fraught with difficulties regarding lack of statistical power and post hoc multiplicity of hypotheses, both in identifying possible geographic differences and pursuing sensible explanations as to why they exist.

Thus, we face an inevitable dilemma in global clinical trial research. On the one hand, we feel obligated to search for potential geographical anomalies in any major international trial. On the other hand, such enquiries rarely (if ever) achieve clarity, and are more prone to generate debate rather than agreed conclusions. This challenge is particularly felt by regulators, especially the FDA, and one might question whether geographic subgroup findings can ever robustly justify label or approval restrictions. Greater confidence regarding results of a pivotal trial in specific regions could be achieved if regulators gave clearer guidance on the required number (percentage) of patients to be recruited in their specific region.
Thus, the moral of this is:

(i) if geographical differences appear to exist in a trial, one would be wise not to believe them at face value; and
(ii) if true geographic differences do exist, one probably will not have enough data to establish their presence;
(iii) potential geographical differences (either intrinsic or extrin- sic) should be carefully considered and identified prospective- ly in the study design;
(iv) if geographical differences are observed and are felt not to be a chance finding, biologically plausible explanations should be investigated further, since geography itself is unlikely to be a sensible explanation for any difference.

Conflict of interest: none declared.

Appendix 1

Statistics
In order to illustrate the difficulties in reliably identifying true geographic differences in treatment effect, we describe the following statistical simulation exercise.

Suppose one conducts a randomized, placebo-controlled trial in five geographic regions, with equal numbers of patients in each region. With 2400 patients, the trial could have 90% power to detect a reduction in primary endpoint incidence from 20 to 15% with $\alpha = 0.05$. Suppose we increase this to 3000 patients in practice.

Now, let us impose the following geographic anomaly. Suppose, in truth, that the 20–15% reduction exists in four regions only, but in the fifth region there is no treatment effect (a true 20% rate exists for both active treatment and placebo). Overall, averaging across regions, the true global effect is therefore reduced to 20 vs. 16%.

From 1000 simulations of such a trial of size 3000 patients (with 600 patients in each region), the following results emerge.

For the overall treatment effect, the median $P$-value comparing treatments across all 1000 simulations is $P = 0.004$. That is, even with one ‘outlier’ region, the trial is still reasonably well-powered to pick up an overall benefit of the new treatment.

For the interaction test used to detect heterogeneity in treatment effect across the five regions, the median $P$-value across all 1000 simulations is 0.24. That is, even though there is substantial true heterogeneity present and the trial is adequately powered to detect an overall effect, the trial is grossly underpowered to detect the geographic outlier.

An alternative post hoc interaction test can compare the ‘outlier’ region with the four other regions combined. This test is dubious to perform since it implies one knows in advance which region is the outlier and in practice one lacks that prior insight. But even this dubious interaction test has only median $P = 0.13$, so power is still lacking.

Table A1 shows how increases in sample size to 6000 and 12 000 patients, respectively, make the overall treatment effect become overwhelmingly significant, and then the interaction test begins to have adequate statistical power to detect the true geographic differences.
heterogeneity. For instance, with 12,000 patients, the median P-value for the overall effect is <0.0000001, and only then can we reliably detect the geographic outlier, with the median P-value for the heterogeneity test being 0.013.

To have 90% power to achieve $P < 0.05$ for the heterogeneity test across five regions requires a trial of over 20,000 patients. That is, to reliably detect this type of regional variation, a trial needs to be at least four or five times larger than is needed to reliably detect the overall treatment effect. This scale of trial effort is very rare in practice, in which case any attempts at exploring geographic variations in treatment effect are irrevocably hampered by insufficient statistical power to do the job properly.

### Table A1 Simulation findings

<table>
<thead>
<tr>
<th>Trial size</th>
<th>3000 patients</th>
<th>6000 patients</th>
<th>12,000 patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall treatment effect</td>
<td>0.004</td>
<td>0.0001</td>
<td>0.00000001</td>
</tr>
<tr>
<td>Interaction test across the five regions</td>
<td>0.24</td>
<td>0.10</td>
<td>0.013</td>
</tr>
<tr>
<td>Post hoc interaction test for outlier vs. the rest</td>
<td>0.13</td>
<td>0.03</td>
<td>0.002</td>
</tr>
</tbody>
</table>

One thousand simulations of a clinical trial is which treatment effect (20 vs. 15%) exists in four regions but is absent (20 vs. 20%) in a fifth ‘outlier’ region. Each P-value in this table is the median one from the 1000 simulations.

### References


