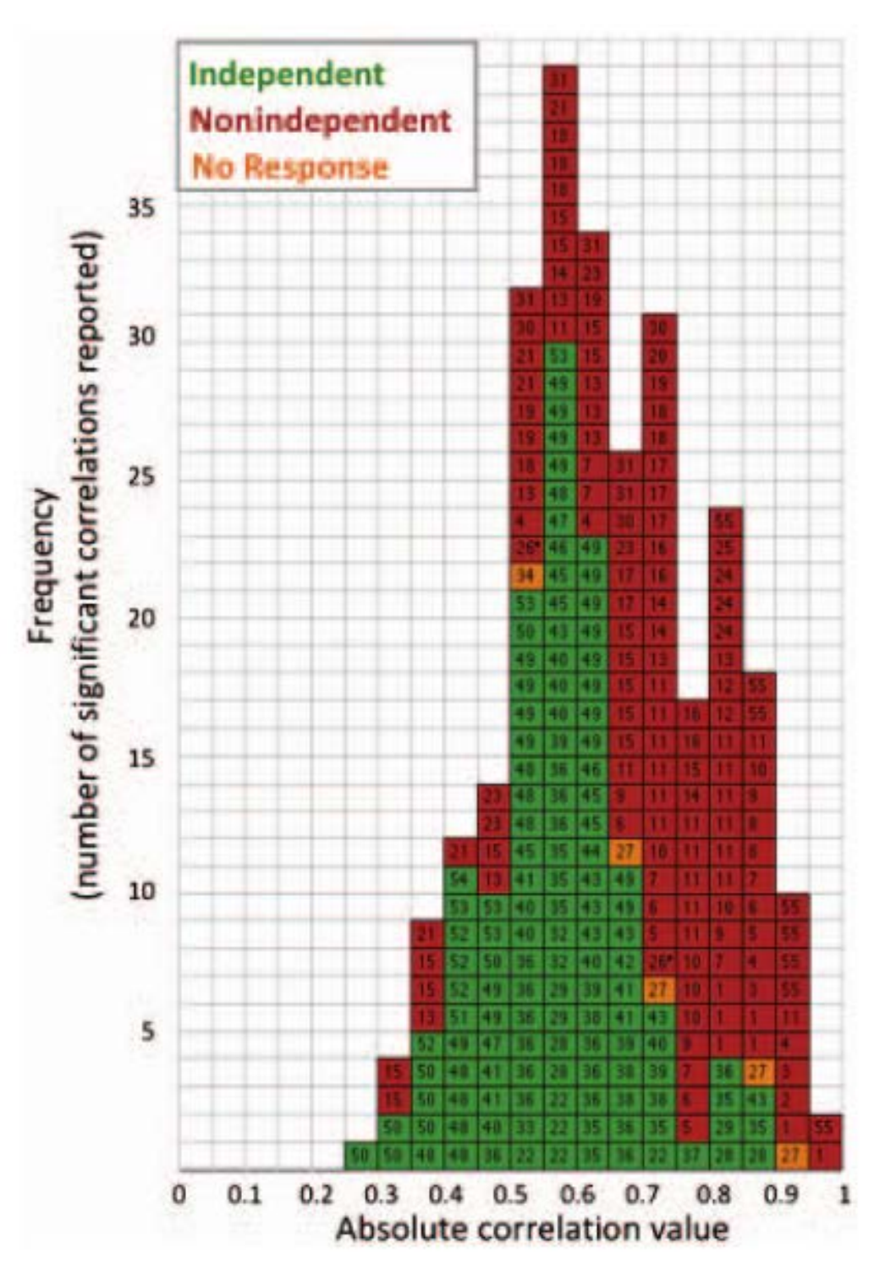
Voodoo-corrected effect sizes without data splitting

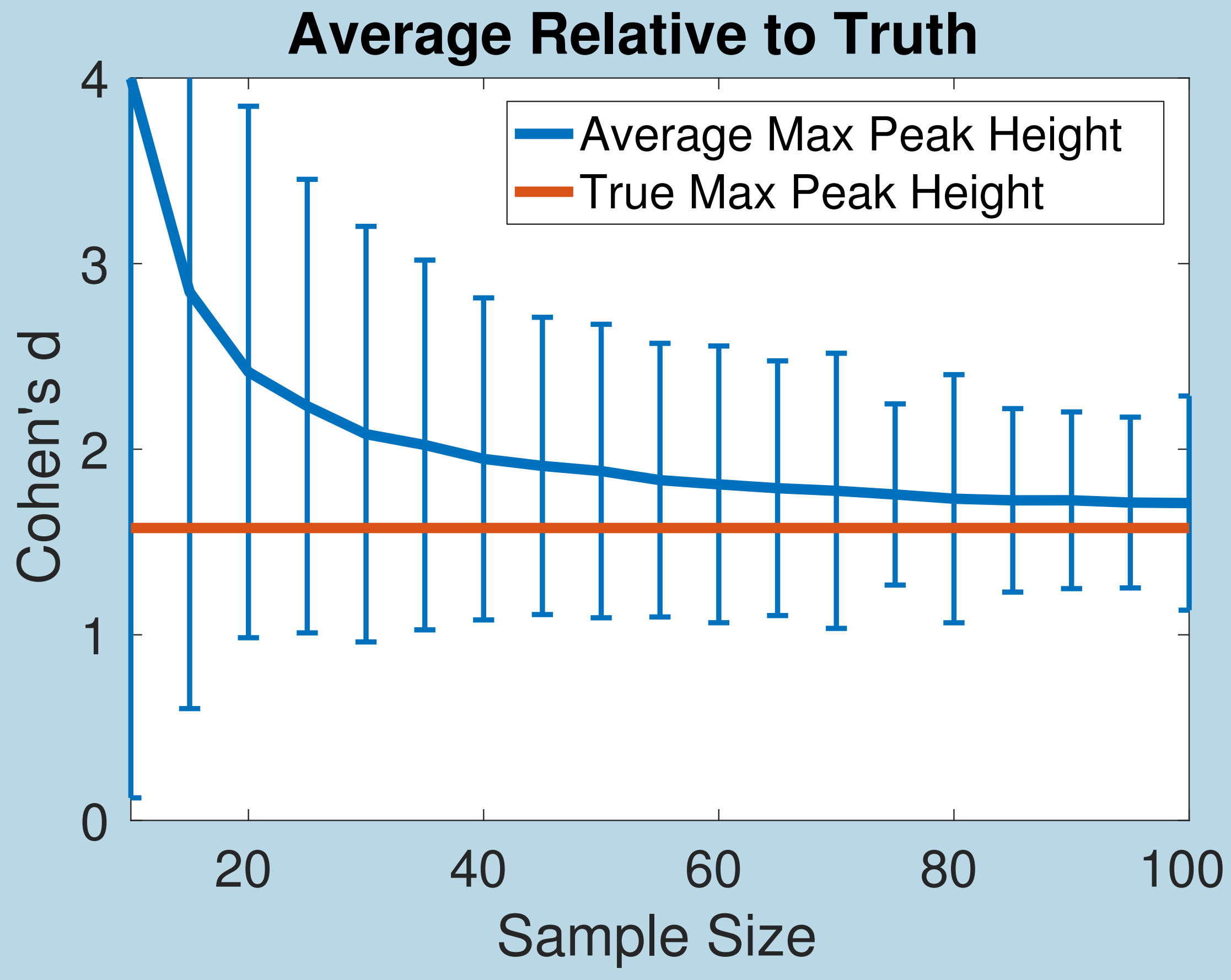
Sam Davenport[1] and Thomas Nichols[2]
[1]University of Oxford, UK [2]Big Data Institute, Oxford, UK

Introduction



Reporting peak effect sizes following a search for significance is an example of the winner's curse problem. Even if an effect is truly present, the magnitude of the effect is over-estimated. This problem was described by the "Voodoo Correlations" paper: Vul et al (2009), that found that "circular" correlations (computed at locations determined from the data, red) usually exceeded non-circular correlations (green). If this problem is addressed at all, the typical solution is **Data-Splitting**, using the first half of the data to find significant regions and the second half of the data to calculate the effect sizes. This produces unbiased estimates, however relative to no splitting, the effect is detected with less power (and spatial accuracy) and the estimate will be more variable. When sample sizes are small, data splitting performs particularly poorly. See Kriegeskorte et al (2010).

Severity of Small Sample Bias



To illustrate the magnitude of the circularity problem curse we compare maximum peak heights as a function of sample size. Using real data (see Methods - Big Data Validation) we computed max peak height for different N, averaged over many datasets, and compare to the true max peak height (from $N = 4000$). The bias is substantial for small N but is nonnegligible even for moderate N. The 95% error bars are based on the 2.5% and 97.5% quantiles for each sample size.

Website and Twitter

@BrainStatsSam, @ten_photos

www.nisox.org

References

[1] E. Vul et al: *Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition,*Perspectives on Psychological Science, 2011.

[2] Kriegeskorte et al: *Circular analysis in systems neuroscience: the dangers of double dipping,* nature neurocience, 2010.

[3] Tan et al: *Selection Bias Correction and Effect Size Estimation under Dependence*, 2014.

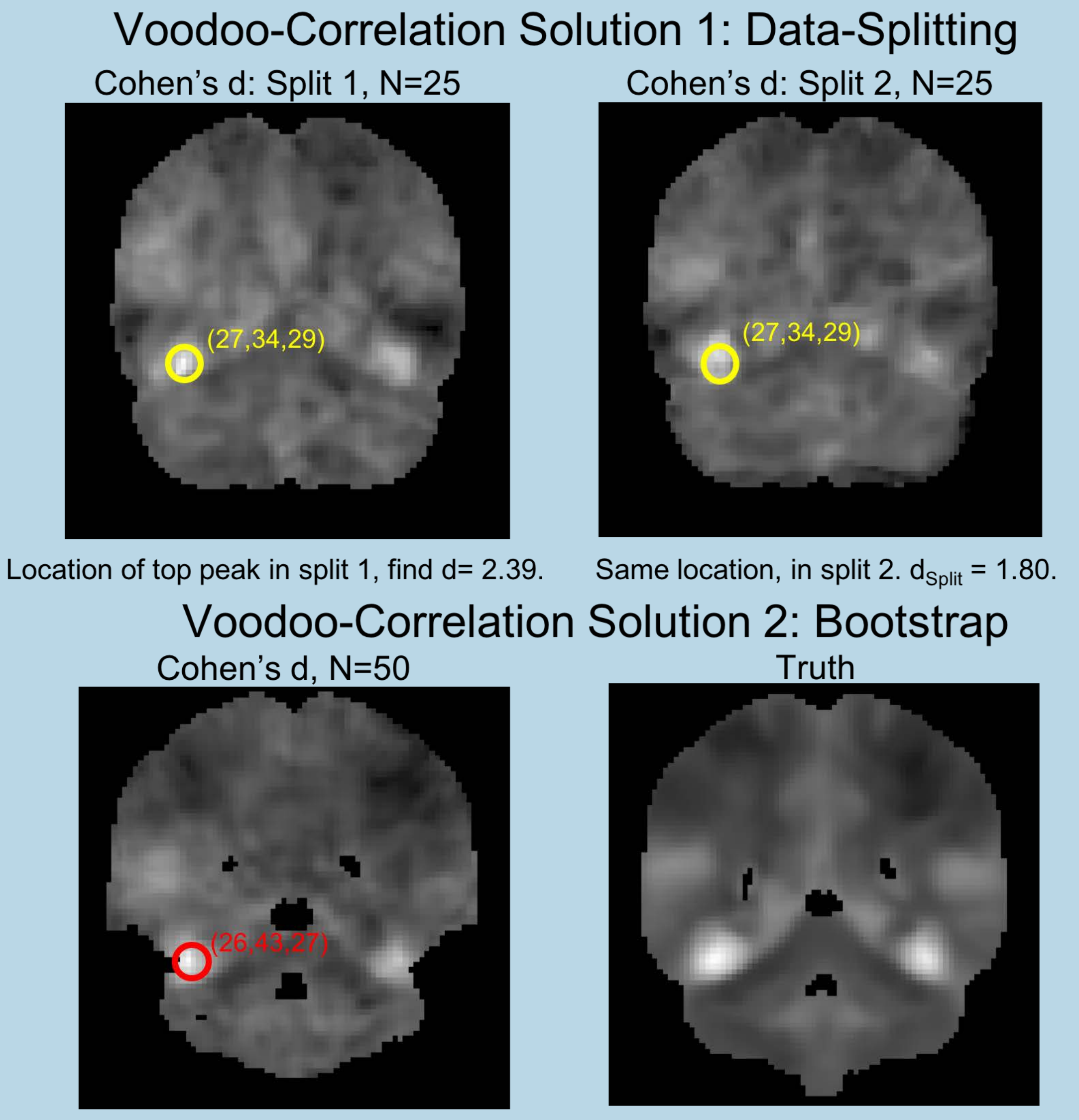
Methods - Bootstrap Bias Correction

We estimate the bias using the non-parametric bootstrap (building on a method based on Tan et al (2014)) and then subtract this bias estimate from the peak effect size to yield a lower bootstrap-corrected estimate of the effect. We compute peak height via Cohen's d, namely:

$$d = T/\sqrt{N}$$

to provide an $N-$independent measure of effect.

In the figure to the right, we have taken an $N = 50$ sample and have illustrated Data-Splitting (top row) and the Bootstrap approach (bottom left). A large-sample truth is shown (bottom right; see details for how it is generated below). The split $N = 25$ samples give noticeably noiser Cohen's d images than the full $N = 50$ image, showing how the bootstrap method can make full use of the data.



Location of top peak in split 1, find d= 2.39. Same location, in split 2. $d_{Split} = 1.80$.

d_{Naive}= 1.99, the bootstrap correction: d_{Boot}= 1.50.

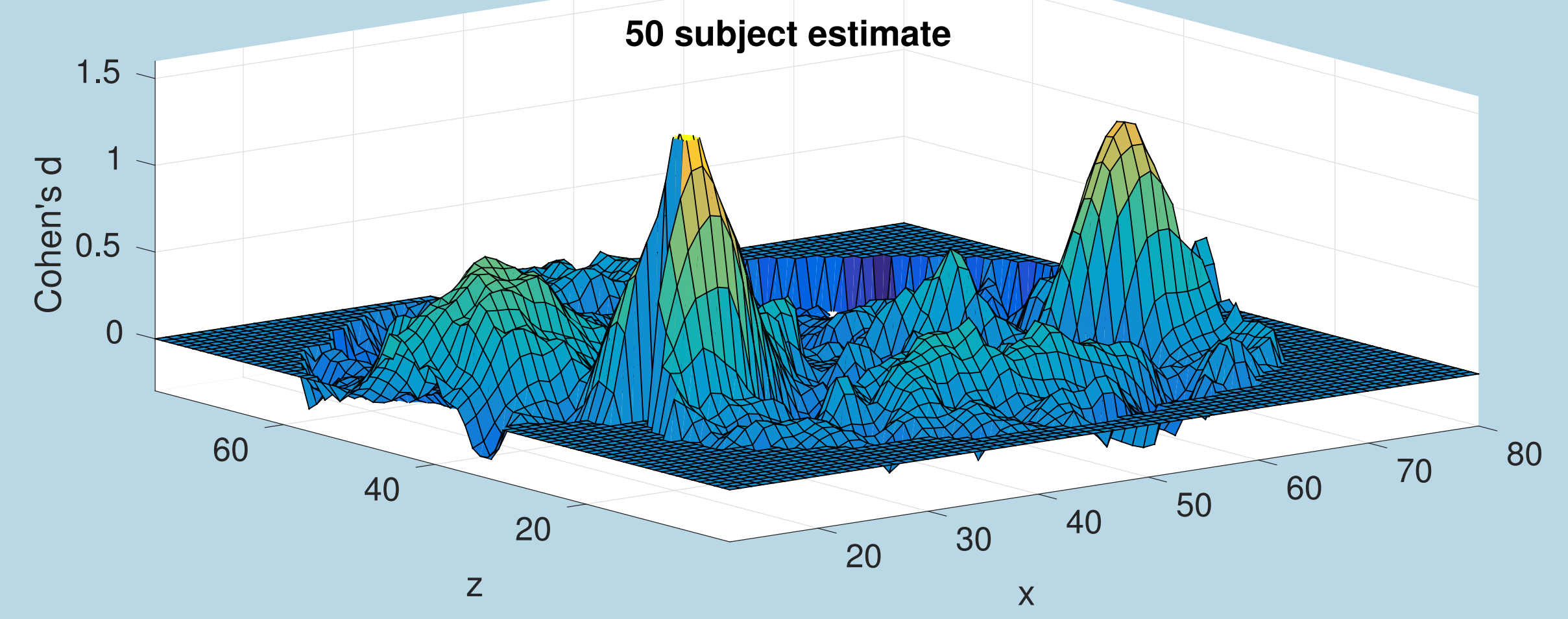
Truth at Data-Split max: 1.23
Truth at Nave Max: 1.39.

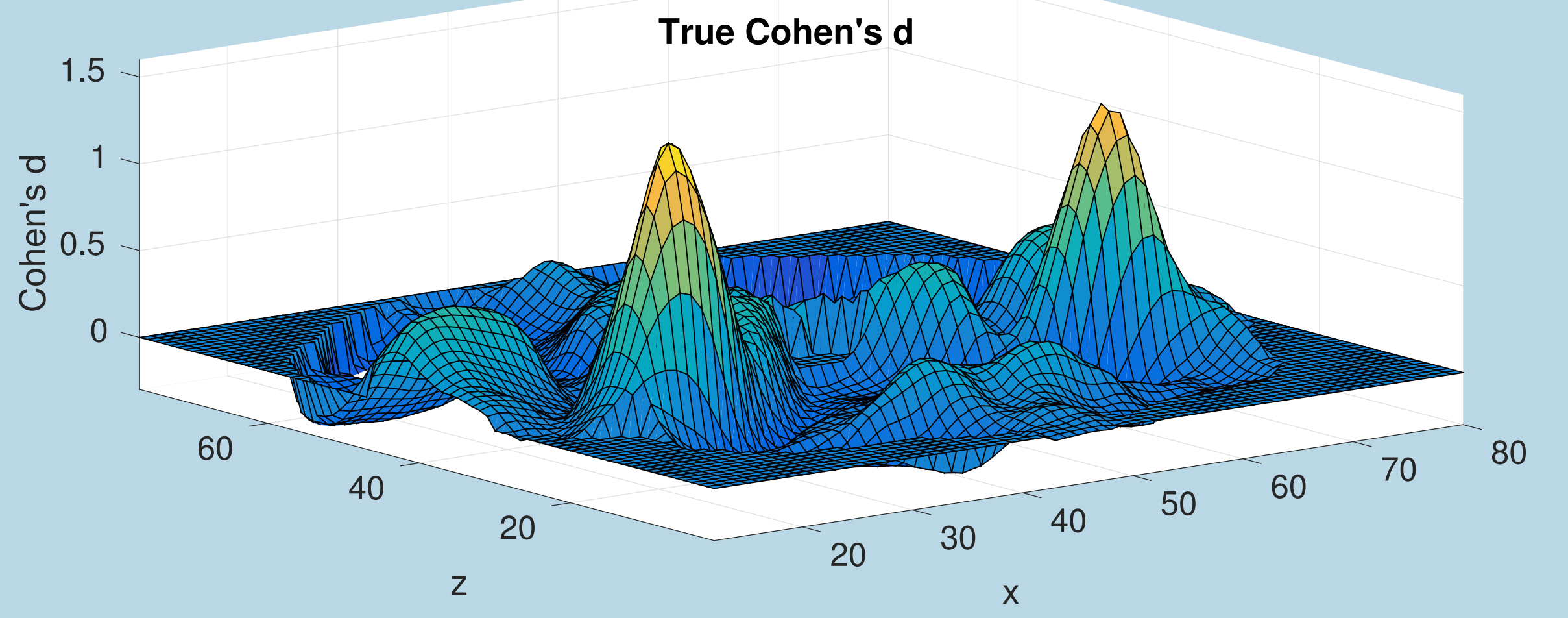
Methods - Big Data Validation

We have used an empirical validation using real data from the UK Biobank with

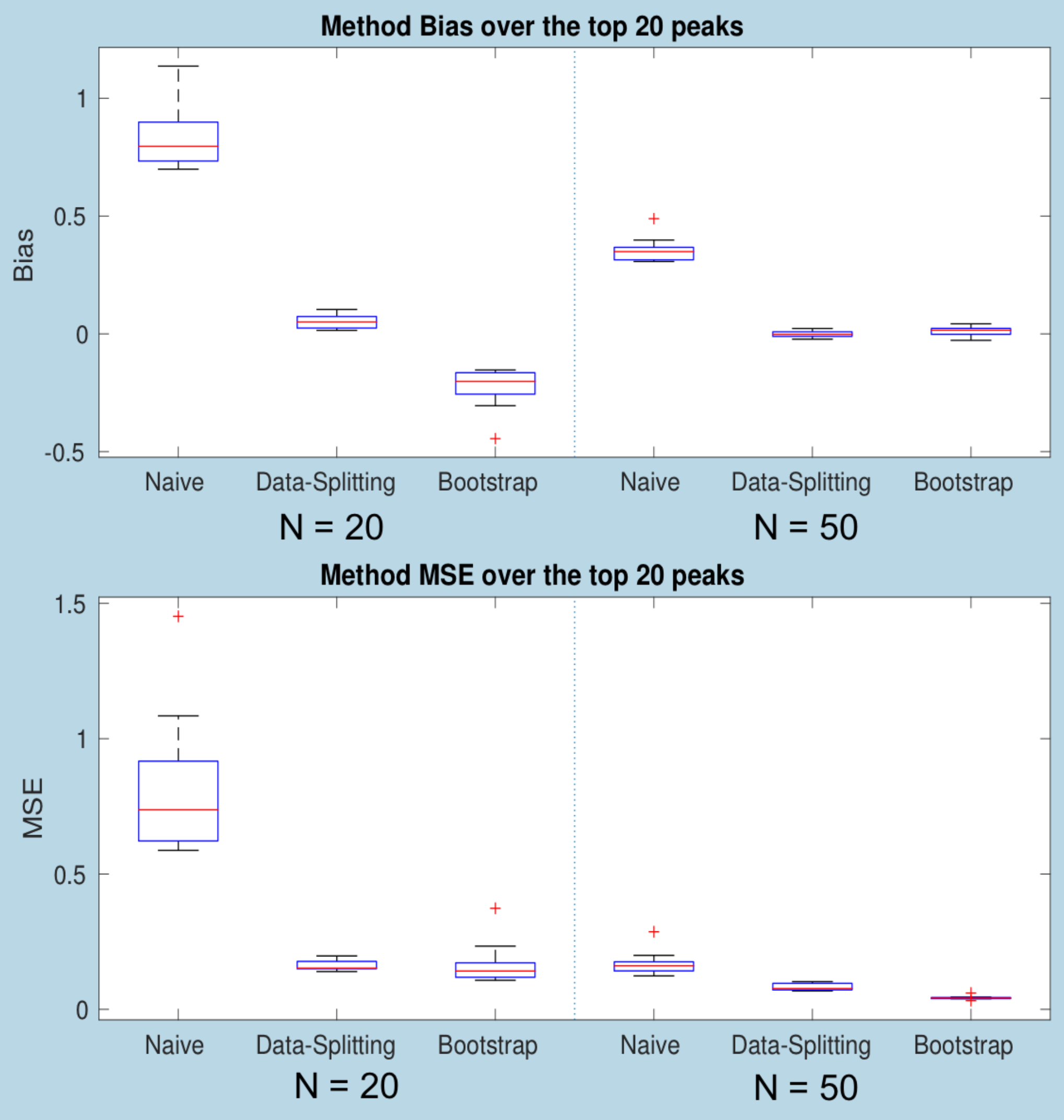
- 8500 subjects and the faces-shape contrast,
- using 4000 subjects to define the truth (figure to the right at the top)
- dividing the remaining available data into 247 groups of $N = 20$ subjects and 98 groups of $N = 50$.

Right we have plotted brain coronal slices of Cohen's d through the truth (bottom panel) and a 50 subject estimate of it (top panel) which match the brain slices above. The top Naive estimate is cutoff as it peaks at 1.99.





Results



In order to compare our methods to existing ones, for each of the 247 groups of 20 subjects (resp 98 and 50) we have found the top 20 peaks in the data for each group and have compared these values relative to the truth (computed using 4000 subjects). The figure to the left shows the box-plots of the average bias (top) and average MSE (bottom) over the top 20 peaks (averaged over all 247/98 groups). The Circular or Naive method has substantially more bias and MSE than either bootstrap or Data-Splitting, though all methods improve with greater sample size. From these plots one can see that the bootstrap and independent splitting methods have low bias relative to the naive method and that the bootstrap method has the lowest MSE. The bootstrap gets better for larger sample sizes as it relies on asymptotic convergence.

Conclusion

Due to the circular inference problem, all of the routinely reported peak statistic values (convertible to Cohen's d), are biased estimates of the true underlying effect. We have proposed a bootstrap based method that dramatically reduces the bias relative to no correction, and has lower MSE relative to the Data-Splitting approach. It has the additional advantage that it uses all of the data to obtain bias-corrected estimates of peak effect sizes and so allows for accurate inference on the location of significant effects (relative to Data-Splitting), as it uses all of the subjects to estimate the locations of significant voxels rather than half of the subjects. Here we have only addressed the one-sample t-test. However it can easily be extended to two-sample t-tests and general regression based settings.