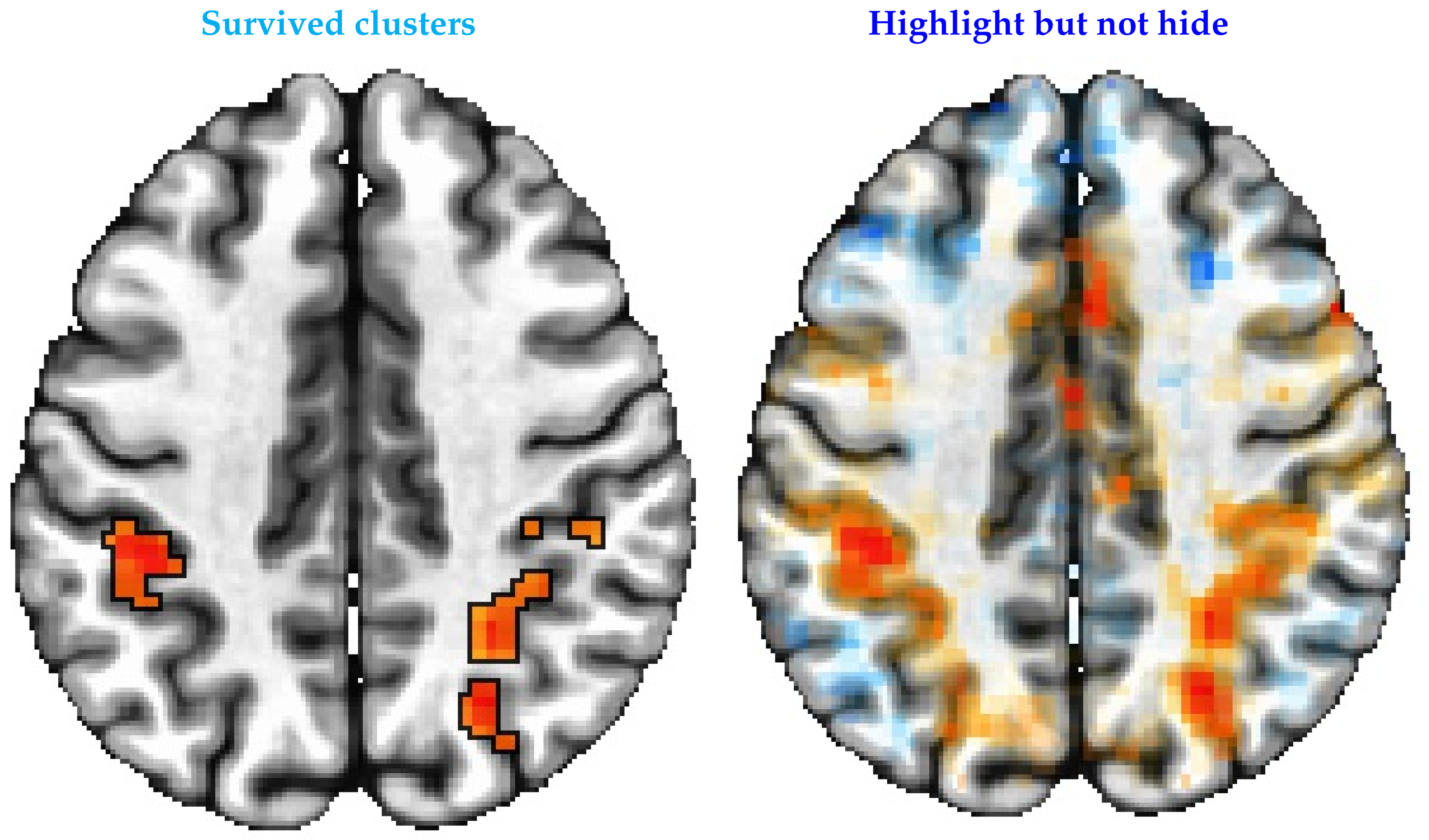
Does the cluster thresholding strategy waste too much information?

Gang Chen[1], Paul Taylor[1], Joel Stoddard[2], Robert Cox[1], Peter Bandettini[3], Luiz Pessoa[4]

1. Scientific and Statistical Computing Core, National Institute of Mental Health, NIH, USA
2. Department of Psychiatry, University of Colorado, USA
3. Section on Functional Imaging Methods, NIMH, National Institutes of Health, USA
4. Department of Psychology, University of Maryland, USA

Correspondence: gangchen@mail.nih.gov

Two types of result presentation in neuroimaging



- Which presentation would you prefer?
  - Left (current practice): rigorously defined clusters per multiple testing adjustment
  - Right (highlight but not hide): strong evidence highlighted with weaker evidence faded
- Questions
  - Which presentation is more informative & realistic?
  - How rigorous/accurate is cluster-level FWE/FPR in the current practice?
  - How is massive univariate modeling associated with excessively conservative inferences?
  - Is artificial dichotomization absolutely necessary in result reporting?
  - Can other info (anatomical structure, prior studies) be used as auxiliary evidence?
  - Is it a good idea to combine left & right in result presentation?

Multiple comparisons: enemy worth fight against?

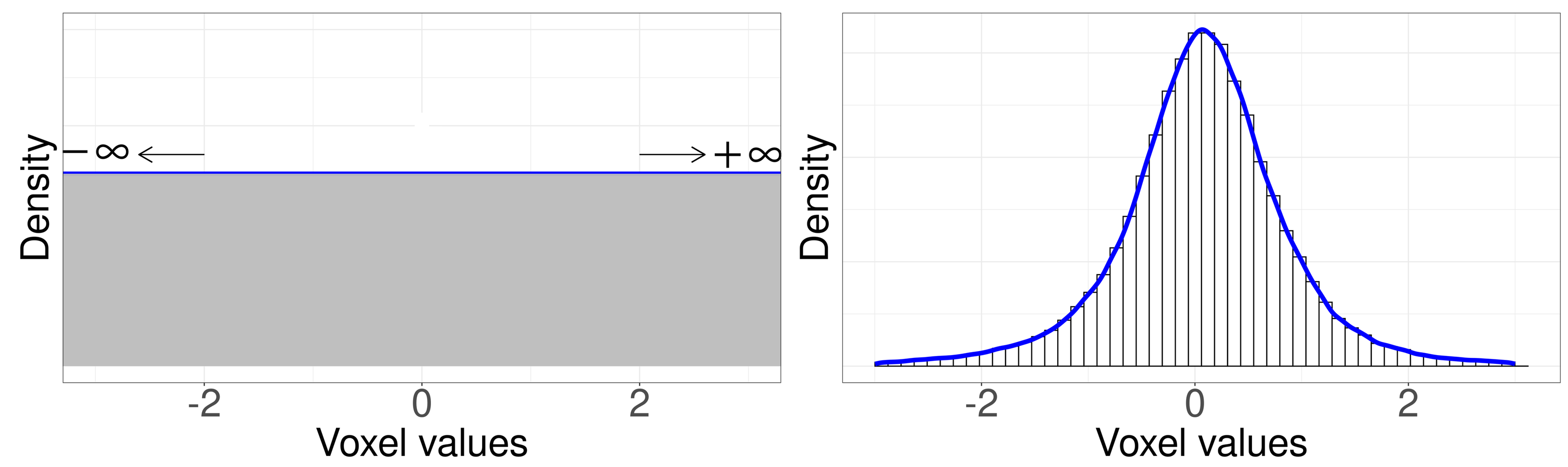
- What is multiplicity, multiple comparisons, or multiple testing?
  - Making a set of statistical inferences that are considered simultaneously
- Situations where multiplicity is a concern
  - Spatial units (voxels, regions, surface node) in FMRI
  - Parameters or effects of interest in a model (GLM, ANOVA, LME)
  - Similar studies for an effect of interest
  - All studies in a journal, a field (neuroimaging) or entire history
- Current solutions to multiplicity: diluting statistical evidence
  - Principle: Bonferroni adjustment - equally diluting/penalizing p-values
  - Variants: trade-off spatial extent against statistical evidence in neuroimaging
    - Monte Carlo simulations; random field theorem; permutations
  - Solving the problem via post-hoc patching-up, not modeling
- Big problems
  - Unbearable penalization: wasting too much info
  - Artificial dichotomization
  - Discrimination against small regions
  - High sensitivity to amount of data (eg, small volume correction)
  - Absence of effect estimation: 1) statistics $\neq$ effect; 2) no adjustment for effect uncertainty
  - Disconnection with anatomy: from statistically-defined cluster to peak voxel
  - Biased selection: exaggeration and incorrect sign; winner's curse, publication bias

Learning from meta-analysis

- NARPS (Botvinik-Nezer et al, 2020): 70 teams independently analyzed same data
  - Teams rigorously followed rules & reported dichotomized results
  - Another multiplicity issue: should team-level p-values be divided by number of teams, 70?
- 2 types of meta-analysis performed by NARPS
  - Weighted average across teams using dichotomized results
    - Logistic model on binarized data ("activated" vs "not activated")
    - Revealed "sizeable variation in the results" across teams
    - Inconsistency: excessive penalty & artitical dichotomization from mass univariate modeling?
  - Weighted average across teams using "unthresholded" statistic values
    - Different conclusion: "yielded a significant consensus"
  - Consistency: which meta-analysis is more reasonable?
    - Similarly, which result representation at beginning of this poster is more reasonable?
- What does meta-analysis mean in result summarization?
  - Statistical evidence: not equal penalization/dilution
  - Weighted summarization: info calibration/regularization/pooling/sharing
  - Importance of full result reporting: reduction of publication bias
  - Each effect, not p-value, is calibrated/adjusted relative to the ensemble
  - Handling multiplicity via modeling, not post hoc patching up
- In handling multiplicity, can we borrow the strategy from meta-analysis?

Recap: information loss in massively univariate analysis

- Same model simultaneously applied to voxels, regions, matrix elements, DTI tracks
  - Multiplicity: as many models as the number of spatial units
  - Local adjustment only: adopted to compensate for multiple testing problem
  - Thersholding: artificially dichotomizing the continuum of statistical evidences



- Implicit but questionable assumption: uniform distribution in $(-\infty, +\infty)$ (Chen et al, 2021)
  - Far from reality: all effects assumed to have same probability of being observed
  - Stance of complete ignorance: excessively heavy penalties, inefficient modeling, poor generalizability, overfitting and compromised predictability
- More accurate characterization of data distribution: centralized density (e.g., Gaussian)

Alternative framework: hierarchical modeling

- Embracing multiplicity: see the forest for the trees
  - Integrative framework that incorporates all spatial units into one model
  - Information partially pooled, regularized and leveraged across all space
  - No multiplicity: one overall high-dimension posterior distribution obtained to infer effects
- Extending meta-analytic strategy
  - Calibrating information, not assuming ignorance (melting, not diluting)
  - Focusing on estimation, not making decision
  - Reporting full results, not hiding ones with weak evidence
  - Quantifying effect & uncertainty, not fighting nullity
  - Emphasizing model quality, not following recipes

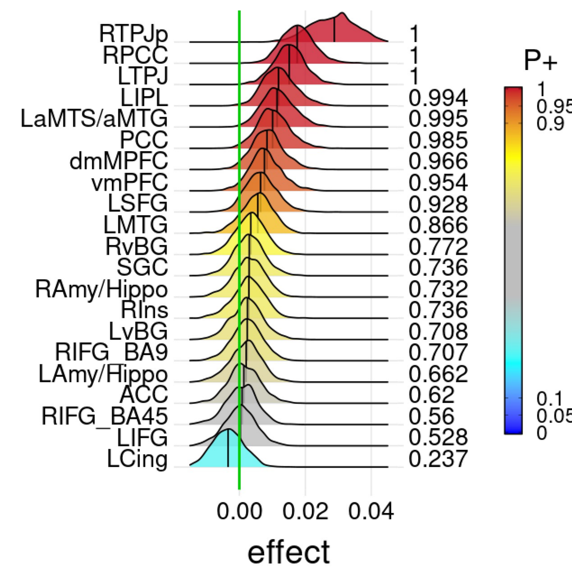
Demo: hierarchical modeling with a dataset

Data information

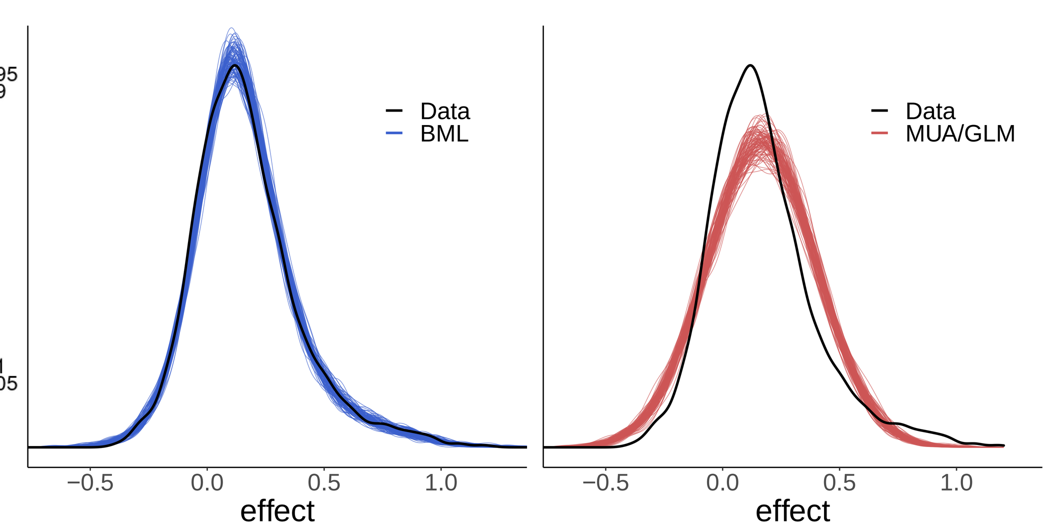
- 124 children watched Inscapes; 21 ROIs
- Explanatory variable: behavioral measure of overall theory of mind ability

Model formulations

- Conventional approach: mass univariate GLM with $\approx 200,000$ voxels/GLMs
  $\boldsymbol{y}_i = a_1 + b_i \boldsymbol{x} + \boldsymbol{\epsilon}_i,\ i = 1, 2, ..., 200000$
- Heavy penalty: 2 regions survived multiple testing adjustment
- Hierarchical modeling (Chen et al, 2019)
  - One integrative model for 21 regions
    $y_{rs} \sim \mathcal{N}(\mu_{rs},\ \sigma^2);\ r = 1, 2, ..., 21;\ s = 1, 2, ..., 124;\ \mu_{rs} = a + bx_s + \alpha_r + \beta_r x_s + \theta_s$
  - Notations
    $a,\ b$: effects shared by all subjects/regions
    $\alpha_r,\ \beta_r$: unique effect of region r; θ_s: unique effects of subject s
  - Explicit assumptions: $\theta_s \sim \mathcal{N}(0,\ \tau^2),\ (\alpha_r,\ \beta_r)^T \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{\lambda}_{2\times 2})$.
  - Information sharing, self-regularization, partial pooling, shrinkage
- Region-level inference based on hierarchical modeling
  - Transparency: full result reporting - highlight but not hide
  - Focus: effect estimation rather than decision making decision through dichotomization
  - Model quality: hierarchical model more closely fitting to original data!



Posterior distributions



Model comparisons

- Moral for massively univariate approach
  - Adjustment based on spatial extent (clusters): excessively conservative
  - Better presenting approach: highlight but not hide

Recommendations

- Region-based analysis: hierarchical modeling
- Whole-brain voxelwise analysis: highlight-but-not-hide
- Focusing on model quality
- Quantifying effect & uncertainty, not making decision via dichotomization

Acknowledgments

The research was supported by the NIMH & NINDS Intramural Research Programs of the NIH.

References

Botvinik-Nezer et al, 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88.
Chen et al, 2019. Handling Multiplicity in Neuroimaging through Bayesian Lenses with Multilevel Modeling. Neuroinformatics 17, 515–545.
Chen, et al, 2021. Sources of information waste in neuroimaging: mishandling structures, thinking dichotomously, and over-reducing data. Aperture Neuro. Articles 46.