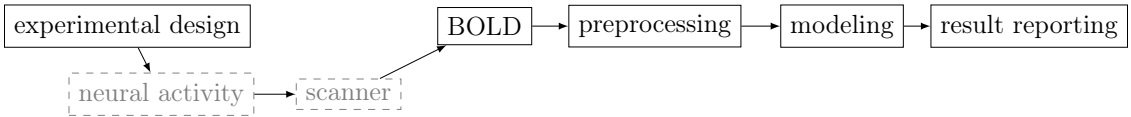
Minimizing Information Waste in FMRI Data Analysis

Gang Chen

Scientific and Statistical Computing Core
National Institute of Mental Health
National Institutes of Health
U.S.A.

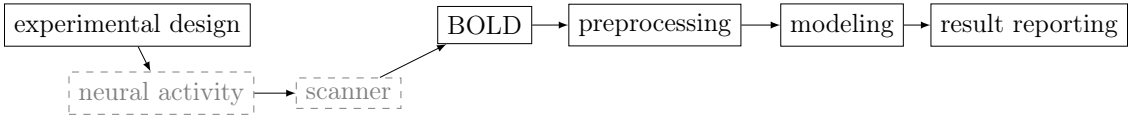
August 20, 2024

Big picture: common fMRI data analysis pipeline



- Data machine: 4 major components
  - ⋆ design: type/quality of data collection
  - ⋆ input: data preprocessing
  - ⋆ device: models
  - ⋆ output: result reporting
- Intertwined components
  - ⋆ output (results): ultimate focus
  - ⋆ streamlined and interdisciplinary
  - ⋆ somewhat disjointed in practice
- Roles of statistics
  - ⋆ statistics rules!
  - ⋆ p-value is everything: colorbars, tables
    - what can be reported
    - which variables considered
- How about auxiliary information?
  - ⋆ previous studies
  - ⋆ data structure/hierarchies
  - ⋆ anatomical structure
  - ⋆ causal relationships

Big picture: common fMRI data analysis pipeline

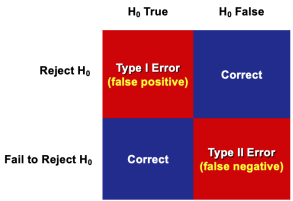


- Experimental design
  - ⋆ type: task, resting, naturalistic
  - ⋆ participants, conditions, trials
  - ⋆ power analysis: sample sizes?
- Input - data quality: preprocessing
  - ⋆ slice timing, motion, spatial alignment, spatial smoothing, temporal scaling
  - ⋆ quality control: data censoring (time points, participants)
  - ⋆ benefits vs harms?
- Device - models: massive univariate
  - ⋆ individual level: regression
  - ⋆ population level: t-test, GLM, AN(C))OVA, LME, ...
  - ⋆ covariate selection, HRF assumption
  - ⋆ challenge: multiple testing problem
- Output - result reporting
  - ⋆ stringency: controlling false positives
  - ⋆ trade-off: info integrity vs digestibility
  - ⋆ thresholding: decision vs estimation?

Traditional framework: null hypothesis significance testing (NHST)

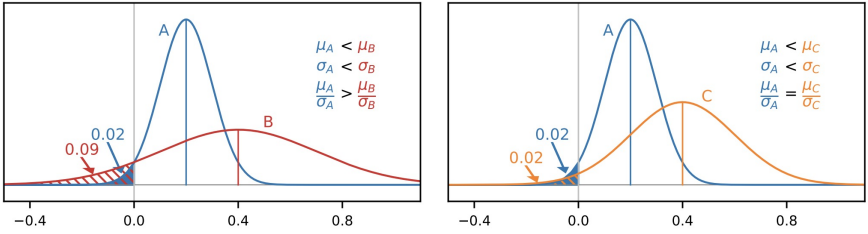
- Null hypothesis (straw man) H_0: zero effect (no involvement, no difference)
  - ⋆ model construction: t-test, regression, GLM, AN(C)OVA, LME, ...
  - ⋆ preset threshold: type I error or significance level α (e.g., magic number 0.05)
  - ⋆ measuring surprise p: conditioning on H_0, how unlikely would real data occur?
  - ⋆ decision-making: gate-keeping process - $p < \alpha$?

- Various problems
  - ⋆ arbitrary: God loves 0.06 nearly as much as 0.05
  - ⋆ fully overlooking type II error
  - ⋆ dichotomization
    - courtroom: innocent until proven guilty
    - scientific investigation: decision-making?
    - sole reliance on statistics: domain knowledge, prior information?



Science is more than just statistics

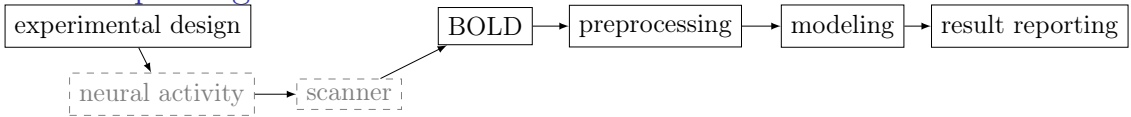
- Pitfalls of solely focusing on statistical evidence
  - ⋆ stronger evidence $\not\Longrightarrow$ larger effect
  - ⋆ equal evidence $\not\Longrightarrow$ equal effect
  - ⋆ speed of light: $p = 0.003$?



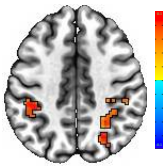
- A different framework
  - ⋆ focus on estimation & uncertainty instead of decision-making
  - ⋆ prior knowledge: causal relationships, previous studies

Chen et al, 2017. Is the statistic value all we should care about in neuroimaging? NeuroImage 147, 952–959

Result reporting



- Reported results: dichotomization
  - ⋆ lack of bilateral symmetry: real?
  - ⋆ border: arbitrary? meaningful?
  - ⋆ part of a region: partial involvement?



- 2 fundamental questions
  - ⋆ research: decision-making process?
  - ⋆ incorporate more information?

- Root problem: modeling approach
  - ⋆ mass univariate analysis
    - same model applied separately: voxel, region, correlation
  - ⋆ multiple testing problem
    - penalty: diluting statistical evidence
    - goal: family-wise error (FWE)
    - method: random field theory, Monte Carlo simulations, permutations
  - ⋆ ritualized procedure
    - surviving clusters at FWE of 0.05
    - critical reviewing process

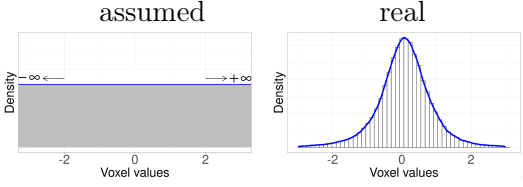
Massive univariate analysis

- Popular modeling approach
  - ⋆ intuitive & computationally economical

$$\begin{aligned} \text{1st voxel: } \boldsymbol{y}_1 &= a_1 + b_1\boldsymbol{x} + \epsilon_1 \\ \text{2nd voxel: } \boldsymbol{y}_2 &= a_2 + b_2\boldsymbol{x} + \epsilon_2 \\ &\dots \\ m\text{-th voxel: } \boldsymbol{y}_m &= a_m + b_m\boldsymbol{x} + \epsilon_m \\ \epsilon_j &\sim \mathcal{N}(0,\ \sigma_j^2); \\ \text{voxel } j &= 1, 2, ..., m. \end{aligned}$$

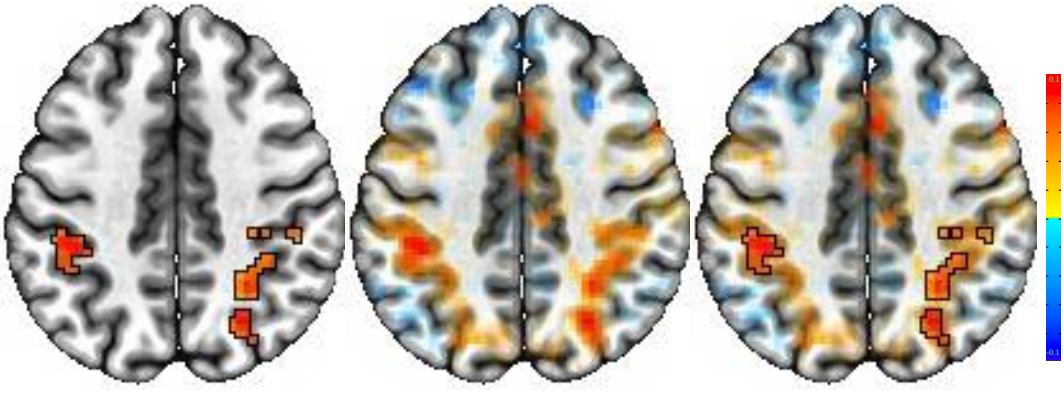
  - ⋆ solutions for multiple testing - penalization (e.g., diluting p-values)
    - random field theory
    - Monte Carlo simulations
    - permutations

- Problems: massive univariate analysis
  - ⋆ implicit assumption: no prior info
  - ⋆ ignoring data hierarchy ⇒ info waste
  - ⋆ band-aid method: adjustments for multiple testing ⇒ excessive penalty
  - ⋆ discrimination against small regions
  - ⋆ ignoring auxiliary info



Chen et al, 2020. Fighting or embracing multiplicity in neuroimaging? neighborhood leverage versus global calibration. NeuroImage 206, 116320

Solution 1: highlight, but don't hide



Taylor et al, 2023. Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. NeuroImage 274, 120138

Solution 2: hierarchical modeling

- Mass univariate approach: many models

$$\begin{aligned}
\text{1st voxel/region: } \boldsymbol{y}_1 &= a_1 + b_1\boldsymbol{x} + \epsilon_1 \\
\text{2nd voxel/region: } \boldsymbol{y}_2 &= a_2 + b_2\boldsymbol{x} + \epsilon_2 \\
&\ldots \\
m\text{-th voxel/region: } \boldsymbol{y}_m &= a_m + b_m\boldsymbol{x} + \epsilon_m \\
\epsilon_j &\sim \mathcal{N}(0,\ \sigma_j^2); \\
\text{voxel/region } j &= 1, 2, ..., m.
\end{aligned}$$

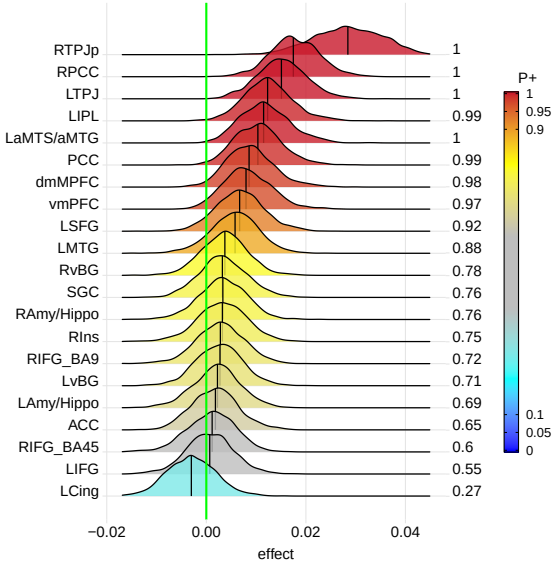
- Hierarchical approach: a single model
  - ⋆ implemented in AFNI program RBA

$$y_{ij} = a + bx_i + \pi_i + \alpha_j + \beta_j x_i + \epsilon_{ij},$$
$$\pi_i \overset{iid}{\sim} \mathcal{N}(0,\ \tau^2);\ (\alpha_j,\ \beta_j)^T \sim \mathcal{N}(0,\ \boldsymbol{\Lambda});\ \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0,\ \sigma^2).$$

Chen et al, 2019. Handling Multiplicity in Neuroimaging through Bayesian Lenses with Multilevel Modeling. Neuroinformatics 17, 515–545

Hierarchical modeling: an example

- Data at population level
  - ⋆ 124 individuals; explanatory variable: behavior measure
  - ⋆ effect of interest: association
- Conventional mass univariate analysis
  - ⋆ 2 clusters survived FWE adjustment based on voxel-level p of 0.001
- Hierarchical modeling
  - ⋆ 21 regions
  - ⋆ using RBA
  - ⋆ full result reporting
  - ⋆ model quality checks: PPC, LOOCV

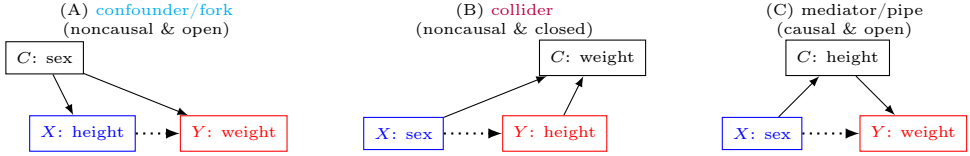


Covariate selection

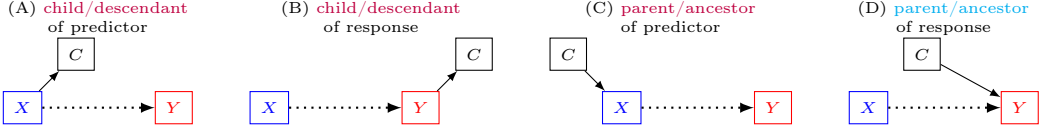
- Statistical modeling
  - ⋆ One model for all effects?
    - step-up/down, statistical metrics (p-values, R^2, information criteria)
  - ⋆ Two goals
    - prediction: forecasting future responses
    - inference: estimating the impact of a predictor on response → causal effects
    - data are amnesic
- An example: data structure for each participant
  - ⋆ response variable: short-term memory (STM)
  - ⋆ predictor: voxel-level gray matter density (GMD)
  - ⋆ 5 covariates
    - 2 between-individual factors: sex, APOE genotype
    - 3 quantitative variables: age, weight, intracranial volume (ICV)
- Questions
  - ⋆ OK to switch predictor and response variable?
  - ⋆ OK to include all covariates?
  - ⋆ are all estimated effects interpretable?
  - ⋆ could more variables have been collected: height, sleep data?

Directed Acyclic Graph (DAG)

- Express prior knowledge or hypothesized relations among variables with graphs
  - ⋆ nodes: variables; arrows: directional influence
  - ⋆ directed acyclic graph (DAG): a common language of graphical representation
  - ⋆ jargon: causal path, front/back door, minimally sufficient set, ...
- 3 basic types



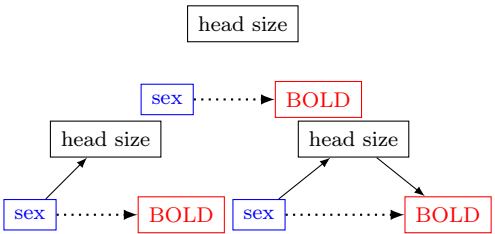
- 4 auxiliary types: covariate influences either predictor or response, but not both



Quiz

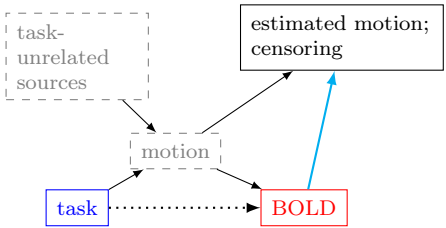
age/site relative to sex/task & BOLD?

head size relative to sex & BOLD



slow drift relative to task & BOLD

head motion relative to task & BOLD

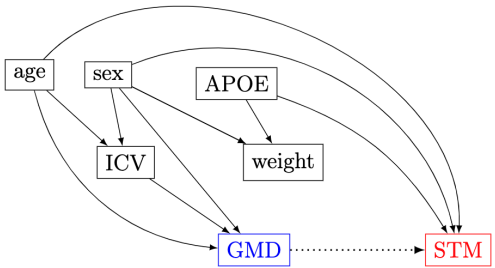


Censoring: data points or participants?

Chen et al, 2024. Through the lens of causal inference: Decisions and pitfalls of covariate selection. Preprint

Revisiting motivating example

- Data structure for each adult participant
  - ⋆ Response variable: short-term memory (STM)
  - ⋆ Predictor: voxel-level gray matter density (GMD)
  - ⋆ 5 covariates
    - 2 between-individual factors: sex, APOE genotype
    - 3 quantitative variables: age, weight, intracranial volume (ICV)



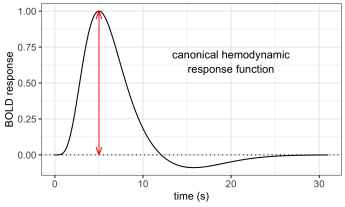
- Addressing four questions
  - ⋆ switch predictor and response variable?
  - ⋆ include all covariates?
  - ⋆ are all estimated effects interpretable?
  - ⋆ could more variables have been collected? height, sleep data?

Summary: variable selection

- DAGs for model selection
  - ⋆ confounder: ✓; collider: ✗; mediator: ⚠
  - ⋆ ancestors/descendants: only condition on ancestors of response
- Suggestions
  - ⋆ drawing DAGs
    - experiment planning & modeling
    - all (including latent) variables
  - ⋆ modeling
    - each effect may require a separate model
    - centering, interactions, nonlinearity
  - ⋆ reporting
    - state effects of interest
    - present DAGs when necessary: transparency
    - avoid listing all estimated effects from a model (table 2 fallacy)
    - avoiding dichotomization: highlight-but-not-hide

BOLD response: standard approach

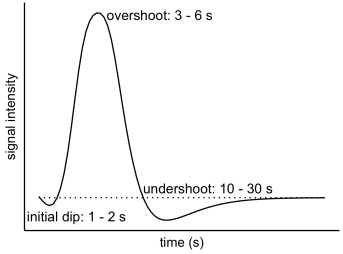
- Canonical: shape-fixed HRF



  - $h(t) = 5.7t^5e^{-t}/\Gamma(6) - 0.95t^{15}e^{-t}/\Gamma(16)$
  - 2 phases: overshoot & undershoot
  - overshoot peaks @ 5s
  - overshoot / overall duration: 12 / 32s
  - undershoot depth: 9% of peak; no initial dip

- Benefit in modeling: widely adopted
  - complexity reduction: 1D $\rightarrow$ 0D (peak height)
  - simplicity: one β per response/condition

- Empirical BOLD response profile

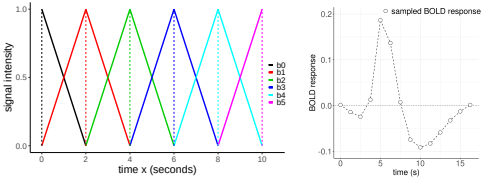


  - 3 phases: initial dip, overshoot & undershoot
  - large variability (eg Handwerker et al 2004)

- Issues with canonical HRF
  - seeing what one wanted to see
  - inflexible: maladaptive to shape variations
  - lost details: peak location, undershoot, ...
  - info loss: inaccuracies & distortion

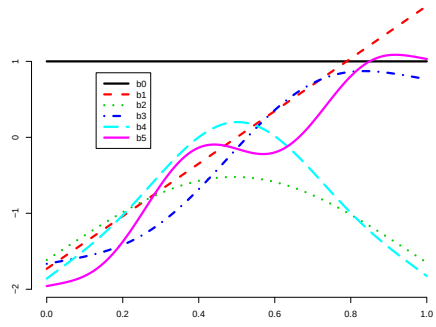
BOLD response: estimation approach

- Estimating HDRs at individual level



  - ⋆ piece-wise linear splines: tents/sticks, FIR
    3dDeconvolve -stim_times 1 stim.1D
    'TENT(2,16,8)'
  - ⋆ estimated HDR: at sampled data points
  - ⋆ shape info: sampled HDR vs 0D (scalar)
  - ⋆ more accurate: data-driven
  - ⋆ weaker assumption: pure morphology vs peak
  - ⋆ challenging for trial-level modeling
  - ⋆ complication: dealing with HDR samples
  - ⋆ sporadically adopted in neuroimaging

- Estimating HDRs at group level: smooth splines

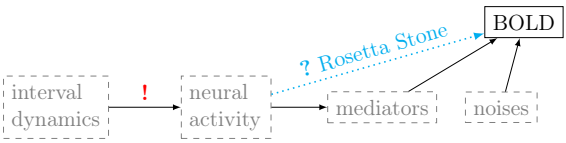


  - ⋆ nonlinear
  - ⋆ smooth: penalization against roughness
  - ⋆ implementation in AFNI: 3dMSS

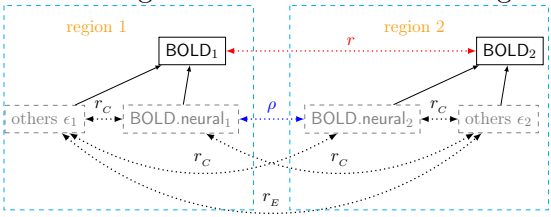
Chen et al, 2023. BOLD Response is more than just magnitude: Improving detection sensitivity through capturing hemodynamic profiles. NeuroImage 277, 120224

Resting-state: how accurate are estimated correlations?

ontological relationships



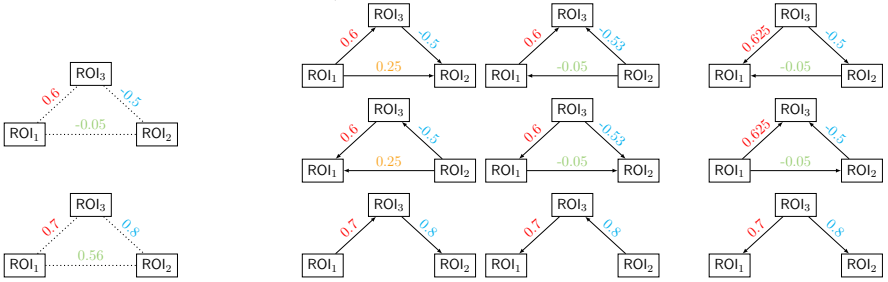
simulating correlation between two regions



- estimating correlations: in the presence of uncorrelated noise
  - ⋆ underestimation (attenuation): Spearman (1904)
- biased estimation due to the presence of mediators & noises
  - ⋆ underestimation: ρ large $\Rightarrow r < \rho$
    - large estimated r rarely seen in literature; BWAS: challenging
  - ⋆ spurious estimation: $\rho = 0 \Rightarrow r > 0$
    - GSR proponents?
  - ⋆ extent of bias: depending on amount of non-neural signal, r_c, r_e
    - denoising wouldn't fully eradicate the issue

resting-state: interpretability of correlation matrix

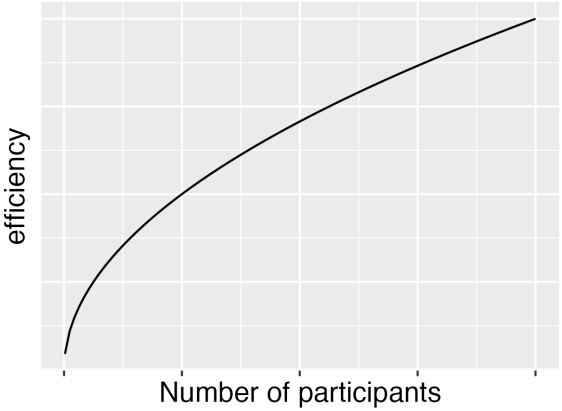
A) Correlations among 3 regions B) Possible causal relationships among 3 regions

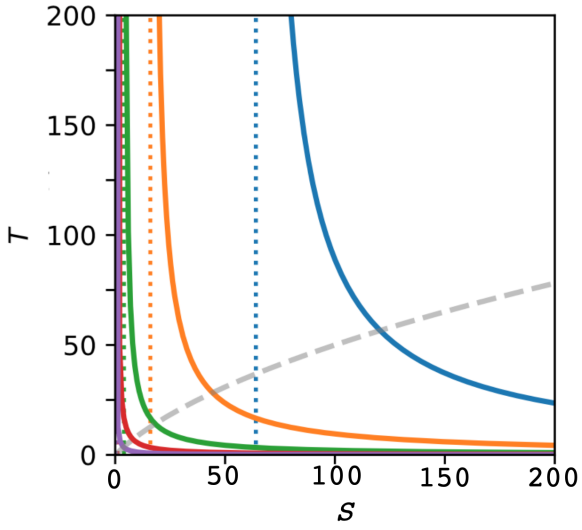


- ambiguities: assuming accurate correlations
  - ⋆ +/-correlation $\not\Rightarrow$ excitatory/inhibitory info flow
  - ⋆ large correlations $\not\Rightarrow$ strong info flow
  - ⋆ small correlations $\not\Rightarrow$ weak info flow

- graph analysis
  - ⋆ nonlinearity, feedback, $>$ 3 ROIs
  - ⋆ thresholding
  - ⋆ topology: hub, centrality, efficiency, rich-club, ...

Role of sample sizes





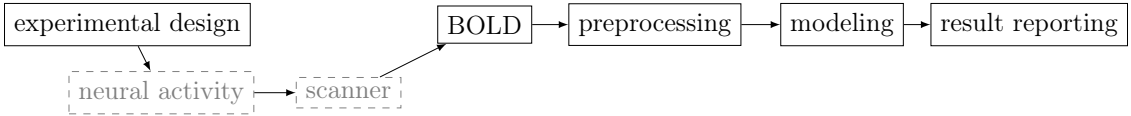
Chen, G, Taylor, PA, Haller, SP, Kircanski, K, Stoddard, J, Pine, DS, Leibenluft, E, Brotman, MA, Cox, RW, 2018. Intraclass correlation: Improved modeling approaches and applications for neuroimaging. Human Brain Mapping 39, 1187–1206.

Sample size considerations

- Difficulty in estimating sample sizes
  - ⋆ effect sizes usually not reported
  - ⋆ results dichotomized at peak voxels
  - ⋆ region-specific: substantial variability across regions
  - ⋆ current power analysis analysis tools
    - solely focusing on participants
    - pacifiers?

- Suggestions
  - ⋆ gather information from literature
  - ⋆ balance trial and participant samples
    - hyperbolic relationship: leveraging between the two in both efficiency and financial cost
  - ⋆ Interactions
    - 2-way interactions: at least a few times more samples than main effects (> 100)
    - 3-way interactions: challenging (> 1000)

Summary: fMRI data analysis pipeline



- Experimental design
  - ⋆ proactively preventing modeling issues
  - ⋆ participants vs trials
  - ⋆ randomization: participants, conditions
  - ⋆ jittering: inter-trial interval
  - ⋆ scanning: space/time resolution
  - ⋆ reducing head motion
  - ⋆ covariate consideration
- Preprocessing
  - ⋆ no one-size-fits-all pipeline
  - ⋆ quality control
  - ⋆ benefits vs harms
- Modeling
  - ⋆ HDR estimation vs canonical HRF
  - ⋆ data hierarchies
  - ⋆ region-based vs voxel-wise
  - ⋆ covariate selection: DAGs
- Result reporting
  - ⋆ highlight, but don't hide
  - ⋆ estimation vs decision
  - ⋆ focus: effect magnitude & uncertainty