Reliability, Power, and Calibration for Multisite MRI Volumetric Studies

Anisha Keshavan[1], Friedmann Paul[2], Mona Beyer[3], Rohit Bakshi[4], Phillip De Jager[4], Massimo Filippi[5], David Hafler[6], Hanne Harbo[3], Stephen Hauser[1], Ludwig Kappos[7], Filippo Martinelli[5], Daniel Pelletier[6], Maria Rocca[5], Till Sprenger[7], William Stern[1], Bernard Uitdehaag[8], Mike Wattjes[8], Howard Weiner[4], Jens Würfel[2], Alyssa Zhu[1], Jorge Oksenberg[1], and Roland Henry[1]

[1]Neurology, UCSF, San Francisco, CA, United States, [2]Charité Universitätsmedizin, Germany, [3]Oslo University Hospital, Norway, [4]Brigham and Women's Hospital, MA, United States, [5]Scientific Institute Ospedale San Raffaele, Italy, [6]Yale University, CT, United States, [7]University Hospital, Basel, Switzerland, [8]Academic Hospital Vrije Universiteit, Netherlands

Target Audience: Researchers planning multisite MRI studies.

Purpose: The benefits of multisite studies come at the cost of scanner/sequence related variation of MRI metrics, prompting many groups to standardize scanners and protocols [1]. However, standardization across multiple sites can be expensive and difficult to implement. Measuring the variability of metrics due to scanner and sequence heterogeneity using traveling subjects enabled us to:

1. Better estimate the power for multisite studies
2. Calibrate metrics when variability between sites is large
3. Optimize processing pipelines and choice of outcome metrics

Ultimately, we aim to understand the conditions under which calibration is necessary for a multisite study with a non-standardized set of scanners and protocols. This is crucial for large genotype/phenotype studies that require a large number of subjects with quantitative MRI phenotypes.

Methods: The magnitude and variability of scaling factors between scanners/protocols for various neuroanatomical volume metrics were measured by acquiring T1-weighted images from 12 subjects (3 Male, 9 Female, ages 24-57) in 9 different 3T scanners (GE, Phillips and Siemens) across Europe and the United States. A neuroradiologist reviewed all images for major artifacts. Standard Freesurfer, FIRST and SIENAX pipelines were run on each site's native T1-weighted protocol. Quality assurance checks on these pipelines failed on T1 spin echo protocols, which were subsequently excluded from the analysis. Test-retest reliability within sites was > 90% for all metrics. Scaling factors between sites were estimated using ordinary least squares, referenced to metrics from the UCSF site.

Theory: A hierarchical linear model for a site *j* and subject *i* was defined as

$$Y_{i,j} = \beta_{0j} + \beta_{1j}X_{i,j} + \beta_{2,j}Z_{i,j} + r_{i,j} \qquad (1)$$

where *X* is a contrast vector, *Z* is a matrix of matched covariates, and *r* is a residual vector with variance σ_0^2. A site-specific coefficient scales the true value of each coefficient, β_{ij}, and the residuals r_{ij}, with scaling factor α_j from a normal distribution with variance σ_α^2. The coefficients were averaged across sites to compute the overall disease effect over *J* sites:

$$\beta_1 = E[\beta_{1j}] = \frac{1}{J}\sum_{j=1}^{J}\beta_{1j} = \frac{\beta_{10}}{J}\sum_{j=1}^{J}\alpha_j = \beta_{10}\alpha_0 \qquad (2)$$

and variance:

$$var[\beta_1] = \frac{1}{J^2}\sum_{j=1}^{J} var[\beta_{1,j}] = \frac{\sigma_0^2\alpha_0^2}{J}\left(\frac{4}{n} + \mathrm{CV}_\alpha^2\left(\frac{4}{n} + \delta^2\right)\right) \qquad (3)$$

where $CV_\alpha = \frac{\alpha_\sigma}{\alpha_0}$ is the coefficient of variability of the scaling factor and $\delta = \frac{\beta_{10}}{\sigma_0}$ is the standardized effect size. In order to test the average disease effect under the null hypothesis that $\beta_1 = 0$, the non-central F distribution is $F\,(1, J-1;\, \lambda)$, where λ is the non-centrality parameter defined as

$$\lambda = \frac{\beta_1^2}{var[\beta_1]} = \frac{J\delta^2}{\frac{4}{n} + CV_\alpha^2\left(\frac{4}{n} + \delta^2\right)} \qquad (4)$$

Results: We found that the coefficient of variability (CV_α) of the scaling factors between sites for most volume metrics derived from MPRAGE volumes were < 5% (Table 1); the non mprage protocol (#7) deviates more from the mprage values. Power curves were generated from eq. 4 (Figure 1).

Figure 1: **A.** Power contours for total number of subjects (nJ) over various effect sizes (d), p= 0.002, CVa = 5%. **B.** # of sites required for effect sizes and # subjects per site (n). **C** effect of CVa on # sites for various effect sizes.

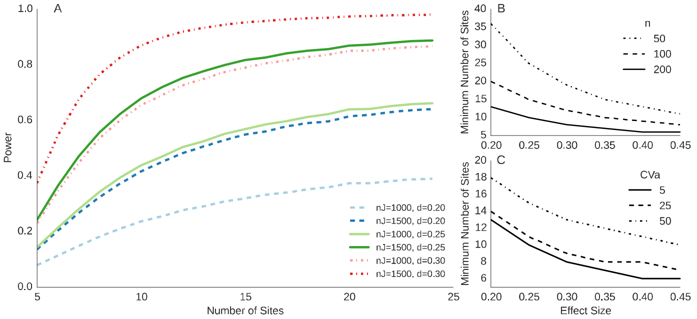

Table 1: Scaling factors to reference sites and coefficient of variability for all vs. MPRAGE sites. Gray matter (GM) White matter (WM) Thalamus (Thal.) Putamen (Put.) Lateral Ventricle (LV)

	Metric				
Protocols	GM	WM	Thal.	Put.	LV
1	0.95	1.00	0.98	0.94	0.97
2	0.92	0.93	0.87	1.00	0.99
3	0.92	0.96	0.97	0.89	0.92
4	0.97	0.96	1.02	0.90	0.96
5	0.99	0.95	0.96	0.99	0.96
6	0.97	1.01	1.00	0.95	0.96
7	*1.31*	*0.59*	*0.86*	*1.15*	*0.94*
8	0.98	0.97	0.98	0.93	1.00
$CV_{\alpha\text{ALL}}$	11.2%	13.2%	5.5%	7.6%	2.5%
$CV_{\alpha\text{MPRAGE}}$	2.8%	2.7%	4.2%	4.3%	2.4%

Discussion: The non-centrality parameter from eq. 4 gives the following limits:

- As $CV_\alpha \to 0, \lambda \to \frac{Jn\delta^2}{4}$ intuitively shows that the total number of subjects (*Jn*) dominates the power equation in the ideal case with no variability between sites.
- As $n \to \infty, \lambda$ is bounded by $\frac{J}{CV_\alpha^2}$ for a non-negligible CV_α. In this case, the number of sites drives the power equation, and it is advantageous to reduce CV_α by calibration with multisite controls.

Options to increase the power of a study are: recruiting more subjects per site, adding a new site, and/or decreasing variability between sites via multisite calibration of scaling factors. Figures 1A and B show that adding a new site was more beneficial than increasing the number of subjects at a fixed number of sites. Figure 1C shows that any imaging metric with CV higher than **25%** (for p=0.002) should be calibrated by multisite controls to increase the power to detect smaller effects. Additionally, we see that the CV for the set of MPRAGE protocols is lower than the CV of the full set, implying that multisite studies may want to restrict their inclusion criteria to protocols with inversion pulses.

Conclusion: For multisite studies using regional volume metrics, the coefficient of variability of scaling factors between sites was low (<5%) for the heterogeneous set of scanners used in this study. Adding a new site generally increases power, however if that site does not use an MPRAGE protocol, the increased variability may decrease the power of the study. Processing T1 spin echo sequences requires specialized pipelines, as they could not be processed by standard methods. Future work includes measuring the scaling factor variability of metrics from other MR modalities (such as diffusion, fMRI, etc.), exploring larger voxels sizes (>1mm) and field strengths, and optimizing morphometric analysis pipelines with the same set of traveling subjects.

References

[1] Tyrone D. Cannon et al. Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis. Human Brain Mapping, 35(5):2424-2434, 2014.