# Reliability of fMRI: Are Group Means Really Representative?

Tynan Stevens[1], Steven Beyea[2], Ryan D'Arcy[3], and David Clarke[1]
[1]*Dalhousie University, Halifax, Nova Scotia, Canada,* [2]*NRC, Halifax, Nova Scotia, Canada,* [3]*Frasier Health Authority, Surrey, British Columbia, Canada*

**Background:** The Rombouts overlap ($R_{overlap}$) coefficient has been widely used for describing reliability of fMRI [1,2]. $R_{overlap}$ is typically calculated from a pair of test-retest fMRI activation maps as the ratio of the common active voxel count to the average number of active voxels. Although it is known that the amount of overlap depends on the threshold used [3,4], no consistent threshold strategy has emerged in studies of $R_{overlap}$ [2]. Moreover, the available literature on the threshold dependence of $R_{overlap}$ is inconsistent, as some authors have demonstrated monotonically decreasing behaviour [2], while others have observed local maxima in $R_{overlap}$ for particular threshold levels [3]. This study investigates $R_{overlap}$ in a series of healthy controls at both the individual and group level. We will show that overlap coefficients are highly variable between individuals. We will also show that interesting features of the $R_{overlap}$ dependence on threshold are lost at the group level.

**Methods:** Eight healthy, right handed volunteers were recruited for this study (4 males, 4 females, 24.4 +/- 3.5 years of age). All eight volunteers were scanned twice with a 4 T scanner (Varian INOVA), for a total of 16 scanning sessions. Test-retest imaging was performed in separate scanning sessions 1-7 days apart. During each session, both structural and functional images were acquired. The structural images were collected with an MP-FLASH sequence (TI=500 ms, TR=10 ms, TE=5 ms, $\alpha$=11°, 256 x 256 matrix, 64 slices, 0.94 x 0.94 x 3 mm voxels). Functional images were collected with a single-shot spiral out sequence (TR = 2 s, TE=15 ms, $\alpha$=90°, 64 x 64 matrix, 22 slices, 3.75 x 3.75 x 5 mm voxels, 0.5 mm gap).

Each participant performed a finger tapping task that utilized a block design, consisting of 20-second alternating blocks of stimulation and rest. Left and right hand ascending/descending thumb-to-digit tapping blocks were interspersed with rest blocks (4 blocks/condition, 9 rest blocks, 170 volumes or 5 minutes and 40 seconds). Pace was fixed at 2 Hz using four circles (for four fingers) to control finger order and timing. Active block order was pseudo-randomized. Stimuli were presented using E-Prime (Psychology Software Tools Inc.), via a projector in the MR console room. Task practice was done before each session to ensure optimal task performance.

Functional MRI analysis was performed using AFNI. Data were motion corrected, segmented (brain/skull), registered (12 parameter affine transformation), and spatially smoothed (Gaussian kernel of 6 mm FWHM) prior to statistical analysis. For statistical analysis, a standard boxcar function was convolved with the default AFNI hemodynamic response. Constant, linear, and quadratic terms were included in the baseline model to account for low frequency drifts. For each unique linear combination of the task regressors, the 3dDeconvolve program was used to produce a t-statistic map. $R_{overlap}$ was calculated from the test-retest scans (see figure 1) for each combination of regressors, over a wide range of analysis thresholds. Reproducibility analysis routines were programmed in python and performed in individual anatomical space.

**Results:** At the group level, $R_{overlap}$ decreases monotonically as the threshold is increased (figure 2). The decrease is slowest along the line $t_1=t_2$. $R_{overlap}$ falls off more rapidly as the difference between the two thresholds is increased (i.e. perpendicular to the $t_1=t_2$ line). At the individual level, the threshold-overlap relationship was more variable, and reliability could often be increased by allowing a non-zero difference between $t_1$ and $t_2$. Figure 2 shows an example of $R_{overlap}$ calculated for a single test-retest image pair. The overlap initially decreases with increasing threshold, but the least rapid decrease occurs for $t_1>t_2$. Furthermore, there is a local maxima in the overlap coefficient at $t_1$=12.5, $t_2$=11. In other test-retest pairs from 0 to 3 distinct local maxima were identifiable. For one dataset three distinct local maxima were observed at ($t_1$,$t_2$)=(4.8,4.8), (12.5,15), (16,19). Local maxima were observed at thresholds as low as $t_1=t_2$=4.5, and as high as $t_1=t_2$=19. Because these maxima do not occur at consistent threshold levels across test-retest pairs, they average out at the group level.

**Discussion & Conclusion:** The observation of local maxima in individual level $R_{overlap}$ plots, and the absence of such a local maximum at the group level, is interesting for several reasons. For instance it may help to explain discrepancies between previous reports of the $R_{overlap}$ dependence on the image thresholds. Whereas the Rombouts et al. report originally showed a local maxima in the overlap coefficient at the group level [3], a monotonically decreasing relationship was reported later by Duncan et al. [4]. We assert that the lack of local maxima in the Duncan report is likely due to averaging at the group level, as our investigation of individual level overlap indicates that these maxima are present int he majority of cases. Interestingly, the Rombouts paper used a normalized measure of brain activity, and this normalization may help to reduce the variability in location of the $R_{overlap}$ maxima (i.e. variability due to overall activation strength).

Furthermore, the presence of local maxima in $R_{overlap}$ at the individual level may provide useful information for localization of brain activity. Thresholds that provide maximal $R_{overlap}$ naturally reveal the most reliably activated brain regions. Given that some datasets contained multiple local maxima, there may be multiple significance levels that can be used on a single dataset to identify multiple reliably activated brain regions. This is especially likely in complex tasks that produce distributed activation patterns with variable activation intensity across multiple brain regions. In conclusion, group mean reliability was not representative of individual reliability behaviour. Thus measuring individual-level reliability routinely is likely the only way to effectively control the reproducibility of fMRI results.

**References:** [1] Rombouts et al. (1997), AJNR 18:1317-1322. [2] Bennett and Miller (2010), ANYAS 1191:133-155. [3] Rombouts et al. (1998), MRI 16(2):105-113. [4] Duncan et al. (2009), 46: 1018-1026.

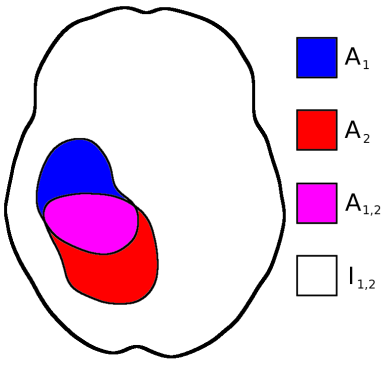


*Figure 1: Illustration of $R_{overlap}$. $R_{overlap}$ is calculated after thresholding the activation maps. $R_{overlap}=2A_{1,2}/[A_1+2A_{1,2}+A_2]$, where $A_{1,2}$ is the volume classified active in both images, and $A_1$, $A_2$ are the volumes classified active on only the first or only the second image respectively.*

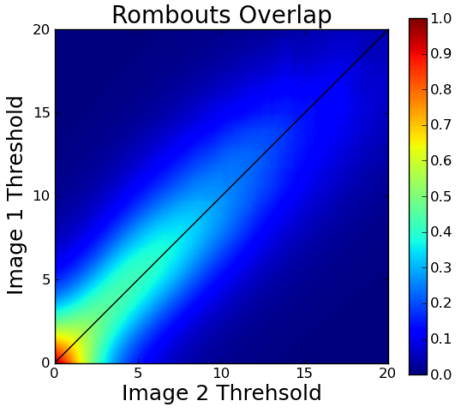


*Figure 2: Group $R_{overlap}$ behaviour as a function of the test and retest image thresholds. At the group level there is 100% overlap when no threshold is applied (the whole brain is active). $R_{overlap}$ decreased monotonically with increasing thresholds (i.e. along the $t_1=t_2$ line), as well as for increasing $|t_1-t_2|$ (i.e. perpendicular to $t_1=t_2$).*

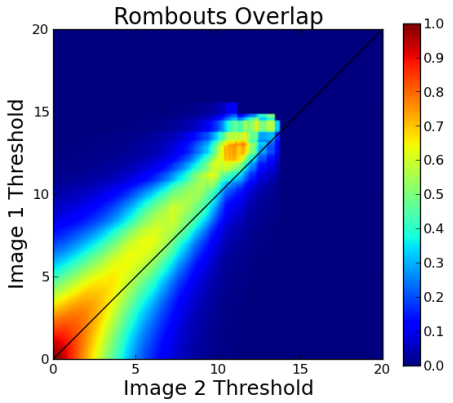


*Figure 3:Individual level $R_{overlap}$ as a function of the test-retest image thresholds. There is a clear local maxima at $t_1$=12.5, $t_2$=11, and in general the greatest overlap is not on the line $t_1=t_2$. Variability between subjects in the location of $R_{overlap}$ maxima leads to their averaging out at the group level.*