Original Research Articles

Understanding variability in brain MRI templates: Optimal sample sizes for representative population averages

Vladimir S. Fonov, Ph.D[1a], D. Louis Collins, Ph.D[1,2]

[1] Department of Neurology & Neurosurgery , McGill University, [2] Department of Biomedical Engineering , McGill University

Keywords: average anatomical templates, variability, sample size, MRI

https://doi.org/10.52294/001c.130075

Aperture Neuro

Vol. 5, 2025

Average anatomical brain templates are routinely used in neuroanatomical and functional studies. Several widely used anatomical models were historically constructed with different algorithms and a limited number of subjects. We performed an experiment to determine the number of subjects required to achieve a stable population average and to estimate variability in individual subjects' registration. We used a random subset of 2000 subjects from the UK Biobank (between 40 and 60 years of age) to generate a "silver standard" population average and then ran a template generation process with a variable number of subjects from 10 to 320, repeating each draw 50 times in a bootstrapping fashion. We compared two methods which are widely used in the literature to generate population averages (ANIMAL and ANTs). Our results showed that 160 subjects are enough to generate a stable population average, and both methods achieve comparable results, with ANTs having advantage over ANIMAL when a smaller number of subjects are available.

INTRODUCTION

Automated image processing in neuroanatomical and functional studies predominantly relies on average anatomical templates as targets for registration to compare data over subjects and between groups. Over the past two decades, numerous anatomical templates have been developed and made accessible to the scientific community. Our group has created a number of human brain templates[1] that have been incorporated into open image analysis tools like FreeSurfer,[2] SPM,[3] and FMRIB[4] to define the target stereotaxic space.[5] Other groups have also created and made available multiple different anatomical templates for the scientific community using variable methods. A recent paper describes TemplateFlow, a web-based repository containing 25 brain templates.[6] Table 1 lists some of the most popular templates.

A common assumption in the creation of average anatomical templates is that the average template is an optimal registration target, allowing the "smallest" parametrization for mapping of each MRI scan in the population into a common coordinate system.[11,16] It has been shown[7,17] that the template which minimizes the average distance (or bias) from all population subjects satisfies these requirements. Such an anatomical template is also called "unbiased", which will be the definition used in this paper.

The number of subjects used to create these average templates often depends on what is available. It remains unclear how many subjects are necessary to capture population variability effectively and produce truly representative, unbiased averages of brain anatomy. Key considerations include the design constraints of templates, such as computational requirements and the availability of the software. Additionally, the methods used for registration and the strategy employed to generate the templates are crucial factors. Other considerations include the intended purpose of templates, which is to facilitate cross-time, cross-subjects, and between-group comparisons. They should also enable accurate registration within studies while minimizing bias. The notion of a minimum deformation target, as discussed by Ashburner,[18] underscores the importance of reducing bias in template construction.

Recently, Yang et al.[19] performed an experiment to determine the optimal number of subjects needed to create a stable population average, using two cohorts: a young Caucasian population from the Human Connectome Project[20] (n=800 subjects, 22–35 years old) and a young Chinese population (n=250 subjects, 19–37 years old). They used random subsamples with a variable number of subjects, ranging from 20 to 400, to generate population averages and then estimated variability of the resulting population templates depending on the number of subjects and race.

[a] Corresponding Author:
vladimir.fonov@mcgill.ca

Table 1. List of several available structural population templates and methods used to create them.

Name	Population	N	Reference	Method
MNI-152	Healthy Adults 18-45 y.o	152	[5] [7]	Variable, most recent one[7]
Colin-27	One healthy adult	1	[8]	N/A
MNI Infant	Healthy infants 0-4.5 y.o	Variable, depending on age 20-50	[9]	[7]
MNI Pediatric	Healthy children 4.5-18 y.o	Variable, depending on age ~ 100	[7]	[7]
ANTs templates IXI,MMRR, NKI, OASIS	Healthy adults 19-90 y.o split into 4 groups	Variable, depending on dataset ~ 30	[10]	[11]
UNC Infant	Neonates 1y.o, 2.y.o	Longitudinal scans of 95 subjects	[12]	[13]
PNC	Young adults 8-22 y.o	393	[14,15]	[15]

In this study, we experimentally investigate the impact of the number of subjects on the creation of unbiased population templates using data from the UK Biobank[21] with a much larger number of subjects. In addition, to ensure generalizability of results, we compare two existing algorithms widely used to generate an unbiased population average: ANIMAL[7] and ANTs,[11] with the aim of determining the number of subjects required for each algorithm to achieve a stable population average.

In a hypothetical ideal situation, if a truly unbiased template were available for the whole study population, we would use it as a “gold standard” reference to compare against templates generated from a subset. Given practical considerations such as computation time, we generated two unbiased templates from 2000 subjects to use as “silver standards” to estimate quality metrics for both methods. Our findings indicate the point at which increasing the number of subjects no longer significantly reduces the variability of results, thereby providing insights into the optimal subject count for template creation.

MATERIALS AND METHODS

MATERIALS

We used MRI scans of the human head available from the UK Biobank.[21] At the time of writing, we had access to the scans of 39677 subjects from version v14940, with mean age of 54.8±7.5 years old. All images were scanned using the same scanner type, a Siemens Skyra 3T, with a 3D MPRAGE sequence (TR = 2000 ms, TE = 2.01 ms, TI = 880 ms, and flip angle = 8°) and 1 mm x 1 mm x 1 mm voxels. All 39677 MRI brain volumes were preprocessed using the pipeline described below. All resulting images were evaluated by automated QC.[22] From the 33725 that passed automatic stereotaxic registration QC, we selected a random subset of 2000 scans (1000 M, 1000 F) of subjects between 40 and 60 years old. See Table 2 below for study population statistics.

ETHICS

Written informed consent was obtained from all UK Biobank participants (see “UK BIOBANK ethics and governance framework”). Research ethics approval was obtained from the North West Multi-centre Research Ethics Committee (for details see https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics).

METHODS

PREPROCESSING

We downloaded raw T1w scans from the UK Biobank repository and pre-processed them using the following steps:

- Intensity non-uniformity correction using N4, parameters: distance 100 mm, iterations: 200x200x200x200,0.
- Histogram-based linear intensity scaling, by matching histogram to that of the MNI-ICBM152 2009c template.
- Linear stereotaxic registration to MNI-ICBM152 2009c space using the *improved bestlinreg* algorithm.[23]
- Automated quality control of the linear registration using DARQ.[22]

REGISTRATION

To ensure generalizability of results, we compared two template building algorithms for the generation of unbiased population templates: The first is based on ANIMAL[7] and the second, on ANTs.[11] Both methods have been used in the literature for construction of several widely used population templates.

In short, both algorithms work by iteratively estimating minimum-deformation templates from a set of MRI scans. The two methods are described below.

Table 2. Study population statistics.

		Female (N=1,000)	Male (N=1,000)
Age (years)	Mean (SD)	51.2 (5.2)	51.3 (5.3)
Height (cm)		163.7 (6.1)	176.9 (6.2)
	Missing	3 (0.3%)	
Weight (kg)	Mean (SD)	68.7 (11.9)	83.3 (12.1)
	Missing	4 (0.4%)	
Waist circumference (cm)	Mean (SD)	80.8 (10.6)	93.1 (9.4)
	Missing	3 (0.3%)	
Ethnic background	White	975 (97.5%)	975 (97.5%)
	Mixed	9 (0.9%)	3 (0.3%)
	Asian or Asian British	5 (0.5%)	9 (0.9%)
	Black or Black British	3 (0.3%)	7 (0.7%)
	Chinese	2 (0.2%)	1 (0.1%)
	Other ethnic group	4 (0.4%)	3 (0.3%)
	Missing	2 (0.2%)	2 (0.2%)
Country of birth	England	837 (83.7%)	842 (84.2%)
	Wales	13 (1.3%)	21 (2.1%)
	Scotland	83 (8.3%)	65 (6.5%)
	Northern Ireland	6 (0.6%)	11 (1.1%)
	Republic of Ireland	7 (0.7%)	4 (0.4%)
	Elsewhere	53 (5.3%)	55 (5.5%)
	Missing	1 (0.1%)	2 (0.2%)
Household income before tax (GBP)	Less than 18,000	90 (9.0%)	60 (6.0%)
	18,000 to 30,999	178 (17.8%)	135 (13.5%)
	31,000 to 51,999	287 (28.7%)	297 (29.7%)
	52,000 to 100,000	279 (27.9%)	318 (31.8%)
	Greater than 100,000	86 (8.6%)	117 (11.7%)
	Missing	80 (8.0%)	73 (7.3%)

ANIMAL-BASED[7]

Uses ANIMAL[24] to perform non-linear registration. Each scan is registered to the current estimate of the template at a given level of detail. The average inverse transform is calculated and concatenated to each individual transformation. The resulting transformation is applied to resample the scan in template space. All warped scans are averaged to create a new unbiased template estimate. These steps form one main iteration. These iterations are repeatedly performed with reducing levels of detail, starting at 16 mm steps for non-linear deformation to address the largest deformations, and ending with 2 mm steps to refine anatomical details. We used normalized cross-correlation as a cost function and the current implementation of the ANIMAL-based template building, available as part of “NIST-MNI image processing pipelines” (https://github.com/niST-MNI/nist_mni_pipelines).

ANTS[11]

Uses ANTs Greedy SyN method.[25] Each scan is registered to the current estimate of the template and warped to match the template. All warped scans are averaged, and average non-linear transformation is calculated. The inverse average non-linear transformation is then calculated and scaled by a gradient step and applied to warp the average to create the new unbiased template estimate. The ANTs template building script applies Laplacian sharpening after averaging individual co-registered scans, resulting in a further increase of the sharpness. See the source code of AverageImages tool from ANTS toolkit, line 147 (https://github.com/ANTsX/ANTs/blob/master/Examples/AverageImages.cxx#L147). This process is repeated iteratively, each time at the highest level of detail (1 mm steps). We used a locally normalized cross-correlation cost function and the script “antsMultivariateTemplateConstruction.sh”, included with the ANTs software package (https://github.com/ANTsX/ANTs), with parameters from.[11]

DIFFERENCES BETWEEN THE TWO METHODS (ANIMAL VS. ANTS) FOR TEMPLATE CREATION

There are several important differences between the two registration methods:

- **Registration method**:
  - ANIMAL uses local Simplex optimization to find deformation vectors optimizing local similarity between corresponding patches. The overall transformation is regularized by applying Gaussian smoothing to these vectors. These steps are repeated for a fixed number of iterations. Inverse deformations are calculated using local Newton optimization. There is no explicit constraint on invertibility of the deformation field or its inverse-consistency.
  - ANTs uses a greedy symmetric normalization algorithm, calculating forward and inverse transformations simultaneously, which are guaranteed to be inversely consistent.
- **Interpolation method**:
  - ANIMAL uses 3rd order b-spline interpolation to resample each MRI before voxel-wise averaging to build the template at each iteration.
  - ANTs uses 2nd order b-spline interpolation.
- **Iterative scheme**:
  - ANIMAL uses a hierarchical scheme, where a minimal deformation template is estimated first at a coarse level of detail, then more refined at each iteration, and so on. In short, at the first main hierarchical step, a 16 mm deformation is estimated for all subjects, and this data is used to generate a 16 mm template average. This process is repeated 4 times to build the final 16 mm template. The final 16 mm template is then used as the target for an 8 mm non-linear deformation estimate for all subjects using the same strategy. This process is repeated until a 2 mm deformation model is created.
  - The ANTs method uses a fixed level of detail and hierarchical scheme at the individual registration level only (e.g., at each iteration, each subject is registered at a 1 mm deformation grid to the current estimate of the unbiased template).
- **Final level of registration detail**:
  - The ANIMAL method stops at a 2 mm step size for the deformation grid.
  - The ANTs method performs registration down to 1 mm steps.

EXPERIMENTAL SETUP

The following steps were repeated with both ANIMAL and ANTS methods:

- We created a single unbiased population average using all 2000 scans to be used as a silver standard template. With such a large number of subjects, we hoped to build a sample mean that would be very close to the population mean. We then estimated a deformation field $T_{i,ss}$ mapping each subject *i* into a common silver standard space (*ss*), and a deformation field $R_{i,ss}$ reverse mapping from common space into each subject's space. These silver standard deformation fields, one from ANIMAL and one from ANTS for each subject, serve as the reference for the analysis below (See Fig. 1.1).
- We randomly chose N=[10, 20, 40, 80, 160, 320] samples to evaluate template creation with different numbers of subjects. We repeated each template creation experiment **50** times for each value of N, drawing new samples (without replacement, without maintaining sex balance from the 2000 scans used to build the silver standard) for each iteration *k*. After each experiment, we again obtained forward and backward transformations: $T_{i,k}$ and $R_{i,k.}$ The rationale for not using sampling with replacement, as commonly done in bootstrapping experiments, is that it is virtually impossible to have two subjects with identical folding pattern, therefore sampling with replacement would create an unrealistic population sample.

We use the following formula to calculate the distance between each transformation and a silver standard transformation:

$D_{i,k} = R_{i,ss} \circ T_{i,k,}$

where ∘ represents concatenation operator. Here $D_{i,k}$ represents a non-linear transformation that can be represented as a dense vector field, defined at each voxel of the silver standard template space, where each vector represents a misregistration "error" between registration parameters of the subject *I* in the experiment *k* from the silver standard (See Fig. 1.1, 1.2). This error can be characterized in several different ways:

1) *Model bias variability*: mean $D_{i,k}$ for each experiment k—showing a mean shift between the silver standard and the subset, followed by standard deviation across all subsets. This metric represents the variability of the k-th average template's overall shape relative to the silver standard.

2) *Individual subject registration variability:* standard deviation of $D_{i,k}$ followed by the average across all subsets. This metric represents variability of the individual transformations to obtain the unbiased template *k*.

3) *Mean deformation value (mDV)*: as introduced in[19] is the mean Manhattan distance of the voxel's displacement distance between silver standard and each template *k.*

4) *Mean absolute logarithmically transformed Jacobian determinant (mALJD)*: as introduced in[19] is a mean of the absolute log-Jacobian of the displacement between the silver standard and each template *k.*

5) *Jacobian determinant variability (JDV)*: is the standard deviation of the Jacobian determinant of $D_{i,k}$ across all subjects in the subset, averaged across all subsets. The Jacobian determinant is a ratio of the local volume change, meaning that value of 0.1 corresponds to the 10% volume change. This metric shows variability of the local Jacobian determinant and can be used to estimate the local effect

size for the power calculation as the denominator in Cohen's d formula in cross-sectional analysis.

OVERLAP METRIC

We used SynthSeg[26] to segment each T1w scan into 32 anatomical regions. We chose this method because it does not depend on registration, and the authors showed high reproducibility and robustness of this method. We used the generalized overlap ratio metric[17] to quantify the accuracy of co-registration of different MRI scans used to create each template, then we averaged (voxel-wise) between experiments for each N, creating *intra-template* overlap maps. We also performed voxel-wise majority voting of each template to generate template segmentation and then calculated voxel-wise overlap ratio between different templates ("*inter-template*" overlap maps). This results in two additional evaluation metrics:

6) *Inter-model generalized Tannimoto coefficient (GTC) overlap*: showing degree of concordance of the segmentation results between templates.[17]

7) *Intra-model average GTC overlap*: average GTC overlap for individual segmentations inside each model that shows the overall degree of consistency of co-registration of subjects with each model.

CONVERGENCE ANALYSIS

Similarly to methods previously proposed[19] we calculated metrics for model bias variability, subject registration variability, mDV and mALJD on a region-of-interest (ROI) basis, and fit a power function to estimate the number of subjects that would be needed to achieve convergence of the template building algorithm, depending on the area of interest that is being studied. We used the following definition of the convergence, similarly to[19]: the number of subjects that are required to achieve below 5% of the slope of the 1st derivative of the power function with respect to the slope for 20 subjects. This means that to achieve a decrease in the metric, the number of subjects that needs to be added to the model is 20 times more than when only 20 subjects are used.

POWER ANALYSIS APPLICATION

Calculating the JDV allows us to perform power analysis for deformation-based morphometry (DBM) in a sense of estimating minimal detectable difference for a hypothetical cross-sectional experiment, with 50/50 split between two groups of subjects, where population average is obtained using either the ANTs or ANIMAL methods, followed by statistical analysis of the Jacobian determinants of the resulting deformation fields. We sampled values of the JDV metric in several anatomical ROIs and fit the same power function as described above to estimate JDV for different values of N. Then, using the sample-size Lehr equation[27] we estimated the minimal detectable difference between two groups of subjects with 50/50 split between groups with α =0.05 and 80% power in a hypothetical cross-sectional experiment:

$$n \approx 16\frac{s^2}{d^2}$$

where *d* is expected difference between mean values of two samples, and *n* is the number of samples in each group.

RESULTS

QUALITATIVE RESULTS

Figures 1.1 and 1.2 show examples of the unbiased averages for N=10, 20, 40, 80, 160, 320 and 2000 scans from the UK Biobank, using ANTs[11] and ANIMAL.[7] While not the true population mean, the N=2000 silver standard is assumed to be quite close. One can see that for N <80, there are evident anatomical differences (see arrows Fig. 1.2) with the silver standard for both ANTs and ANIMAL averages. These images indicate that N < 80 is not enough to be representative of the population average.

Figure 1.3 shows an example of the inter-model variability when a small number of scans are used to calculate the averages. Again, arrows in the figure draw attention to the differences.

QUANTITATIVE RESULTS

VOXEL-LEVEL QUANTITATIVE RESULTS

Figure 2.1 shows the model bias variability and subject registration variability on the voxel level. As expected, there is higher variability at the cortex compared to the deep brain, which decreases with increasing N. The positional variability (i.e., the mismatch with the N=2000 silver standard depending on the anatomical ROI) is minimized using 80 or 160 subjects when using ANTs. The ANIMAL-based unbiased average requires N=160 to achieve a similar result. Interestingly, individual positional variability (i.e., the anatomical variability between subjects) stays roughly the same for all N of both methods.

Figure 2.2 shows the mDV and mALJD. In general, the mDV reinforces the model bias variability, with the difference being the use of Manhattan distance instead of Euclidean distance for characterizing local variability of shape. As such, the same comment as above applies here. The mALJD reflects a different aspect of the model variability: rather than showing displacement in mm, it represents variability of the local volumes. Surprisingly, even though the mDV metric demonstrates a noticeable difference between ANIMAL and ANTs, the mALJD metric shows that there is a little difference between the two methods, although for the small number of subjects (below 20), ANTs seem to produce slightly less variable results.

Figure 2.3 shows the average Inter-Model GTC overlap and average Intra-Model GTC overlap, also on a voxel level. These measures show how well segmented anatomical structures co-register together. As expected, there is higher variability at the cortex compared to the deep brain, and this variability decreases with increasing N.

Figure 2.4 shows the JDV at voxel level. This parameter can be used as the denominator in Cohen's d effect size for-

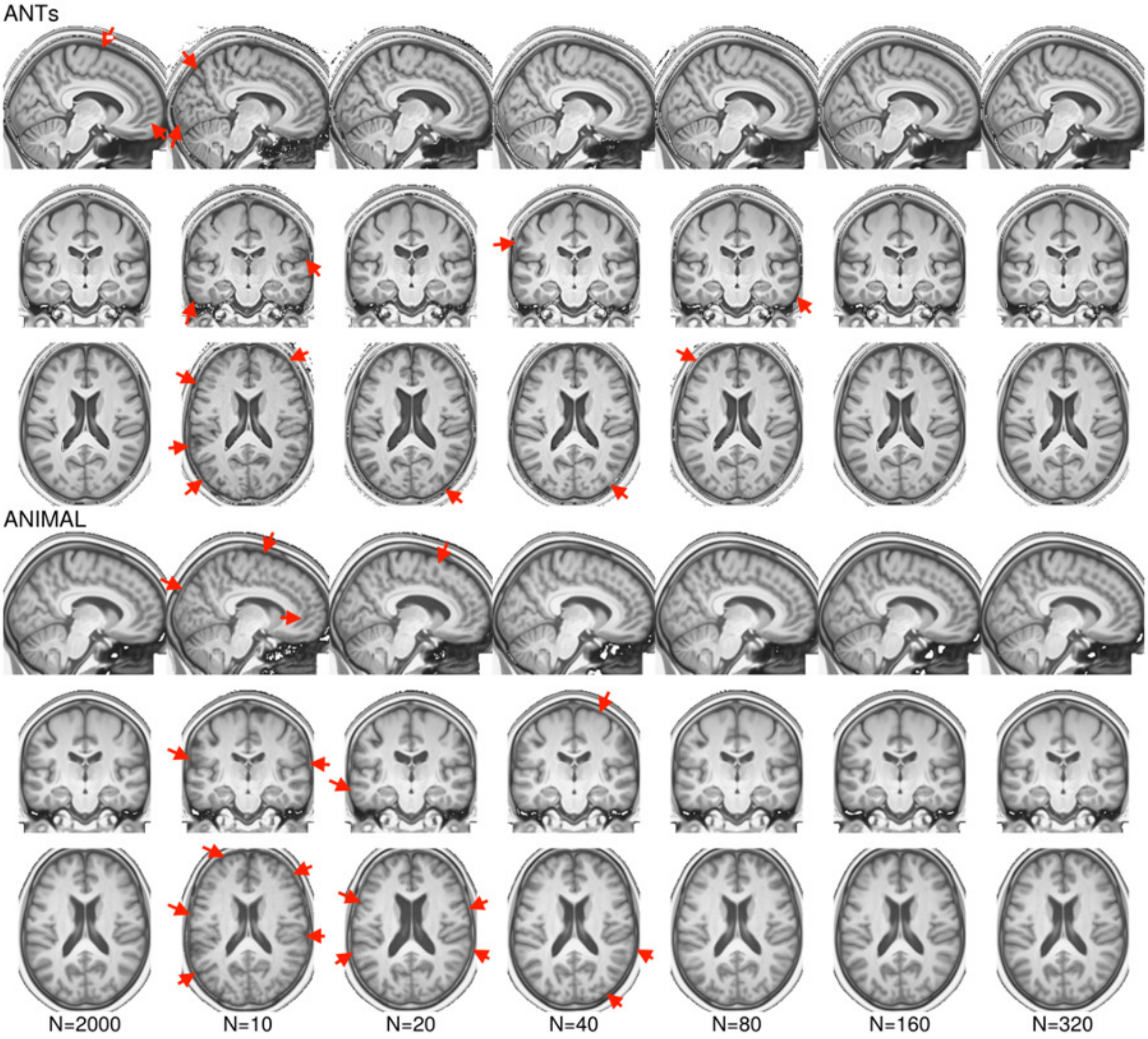

Figure 1.1. Unbiased averages for N=10, 20, 40, 80, 160, 320 and 2000 UKBB scans using ANTs (top row) and ANIMAL (bottom row). Arrows draw attention to anatomical differences with the silver standard.

mula for the power calculation in hypothetical cross-sectional analyses.

CONVERGENCE ANALYSIS RESULT AND REGION OF INTEREST LEVEL QUANTITATIVE RESULTS

Results of the convergence analysis are shown on Figure 3.1 and Figure 3.2, where the fitted functions and red numbers indicate the estimated number needed for convergence. Adding more subjects yields diminishing improvements in model stability. The number of subjects required for model bias variability are similar between ANTs and ANIMAL methods. However, more subjects are required (roughly 2x) for ANIMAL to achieve equivalent subject registration variability. The mDV and mALJD are very similar between methods.

POWER ANALYSIS RESULTS

Figure 3.4 shows the results of the power analysis estimation of the minimal detectable difference between two groups of subjects with 50/50 split between groups with α =0.05 and 80% power in a hypothetical cross-sectional experiment. Note that the y-axis is shown on a log-scale and units are the percent of the local difference between two groups of subjects.

DISCUSSION

In this paper, we have built the largest (N=2000) unbiased average template MRI brain model to use as a silver standard of the average brain shape. We used data from the UK Biobank—high quality, high resolution, high contrast T1w MRI data from subjects with a broad age range, acquired on identical scanners that enabled us to focus on anatomical differences and not scanner differences. Our compar-

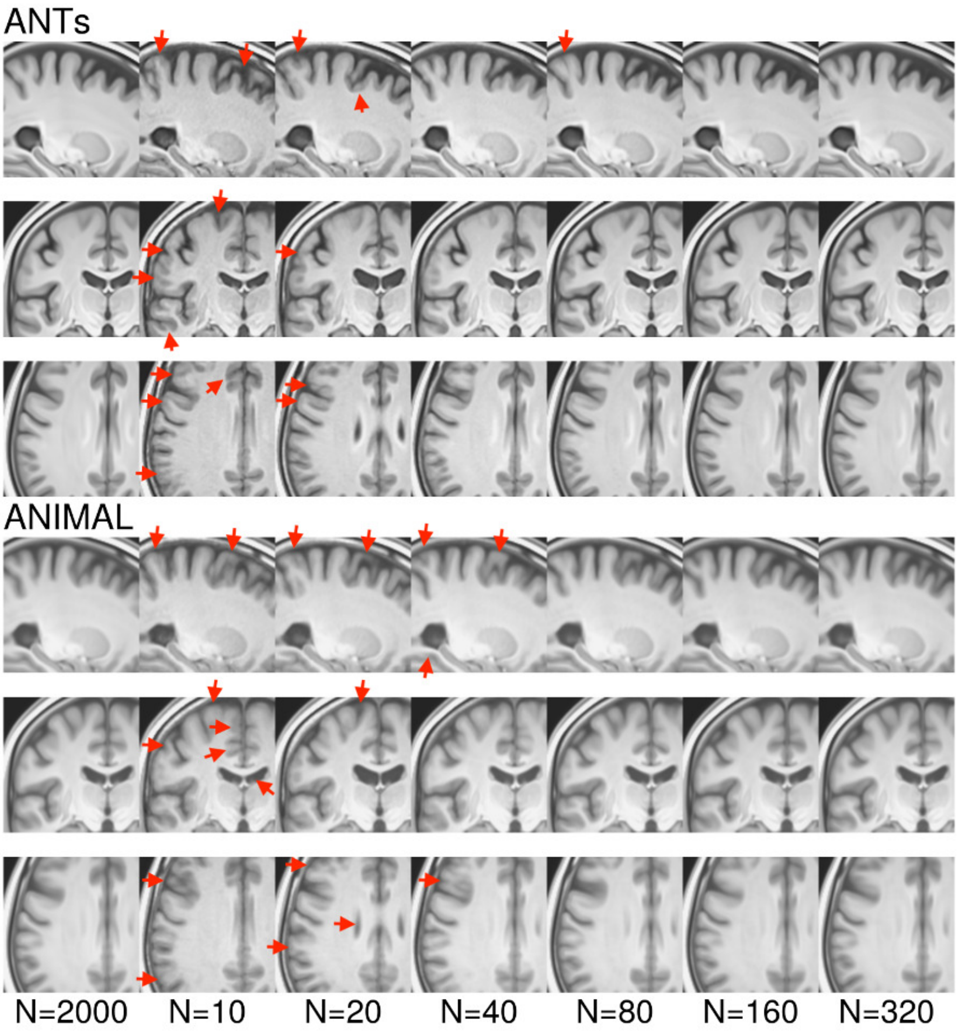

Figure 1.2. Unbiased averages for N=10, 20, 40, 80, 160, 320 and 2000 UKBB scans using ANTs (top row) and ANIMAL (bottom row). Closeup view. Arrows draw attention to anatomical differences with the silver standard. Recall that ANTs template building uses a Laplacian sharpening step. This might explain the higher apparent edge contrast in the ANTs templates.

ative experiments demonstrate that both methods, ANTs and ANIMAL, can successfully build unbiased average templates.

The quantitative results show that as far as the shape of the average template is concerned, the ANTs method achieves smaller variability of the overall shape of the template for a given number of subjects. This is particularly noticeable in the cortical regions, where inter-subject variability is reduced by a factor of two with the ANTs method. Figure 3.1 shows that for inter-subject registration variability estimated in the cortical grey matter (GM) ROI, the ANTs method converges after 240 subjects with remaining variability of roughly 1 mm, whereas the ANIMAL method converges after more than 400 subjects, with residual variability of approximately 2 mm. This is perhaps due to two reasons. First, the ANTs registration algorithm runs to 1 mm deformation field step size, whereas the ANIMAL method stops at a rougher resolution of 2 mm. Second, ANTs uses symmetric normalization of the deformation fields, while ANIMAL uses simple Gaussian blurring of the deformation fields for regularization.

Apart from the geometric measurements derived from the registration parameters of the silver standard, we have tried to estimate goodness of co-registration by measuring degree of agreement of segmentations between models. We use the overlap metric as another proxy of the goodness of the model.

The inter-model overlap metric plotted in Figure 3.2 shows that with an increased number of subjects, the overlap gradually reduces in the cortical GM ROI, for both ANTs and ANIMAL, possibly indicating that **neither of the proposed methods is able to fully capture the anatomical variability of the cortical folding pattern**. Another potential possibility is that voxel-based non-linear registration methods like ANTs and ANIMAL are not sufficient to co-register cortical gyri and sulci and surface-based methods should be used instead when accurate cortical co-registration is needed.

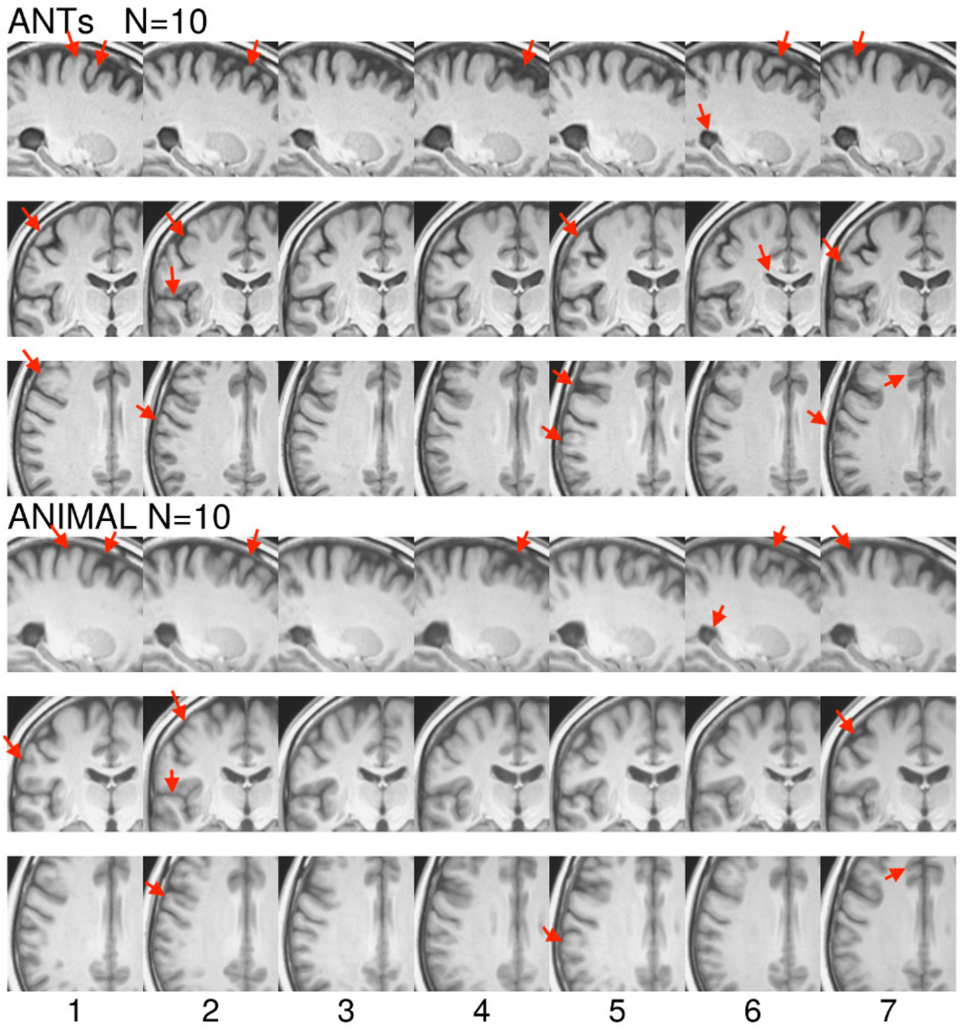

Figure 1.3. Individual examples of unbiased averages for N=10 UKBB scans using ANTs (top row) and ANIMAL (bottom row). Closeup view.

Interestingly, both methods converge to a stable population average when measured with residual model bias variability of ~ 0.25 mm and mDV of ~ 1 mm after roughly 160 iterations in all regions of the brain (see model bias variability in Figure 3.1, and mDV in Figure 3.3). The only exception is for the brainstem and cerebellar white matter (WM), where ANTs surprisingly needs more than 200 subjects to converge to a stable population average.

The JDV metric (Figure 2.4) can be used to estimate the smallest possible effect size that can be captured reliably when non-linear registration methods are used for DBM experiments to determine the minimal effect size that can be reliably captured. This metric could be used to help estimate the required number of subjects in a power calculation for pharmaceutical trials that target specific brain regions.

Our findings are consistent with those published in [19], which showed that 200 subjects are enough to achieve a stable population average, although using a smaller population pool and using a single template building algorithm (ANTs).

Our power analysis estimations show one of the practical applications of the result of this study: estimating minimally detectable change for statistically significant result at α =0.05 with 80% power in a hypothetical DBM experiment, or alternatively one can estimate required number of subjects for an expected difference between two populations, depending on the anatomical location. Results show that ANTs has a slight advantage over ANIMAL with regard to the number of subjects required to detect a given difference, especially in the cortical GM. Also, clearly visible is that a smaller number of subjects would be needed in deep GM, brainstem, and cerebellar WM structures than anywhere else.

Our paper is not without limitations. First, while we only compared ANTs and ANIMAL algorithms, we would expect similar results with other unbiased average template creation strategies. Second, we ran ANIMAL using default parameters that estimated the final deformation grid at a 2 mm spacing, potentially limiting its performance in comparison to the ANTs deformation grid at 1 mm spacing. Third, the ANTs method includes a Laplacian sharpening step that increases the GM:WM contrast in the average

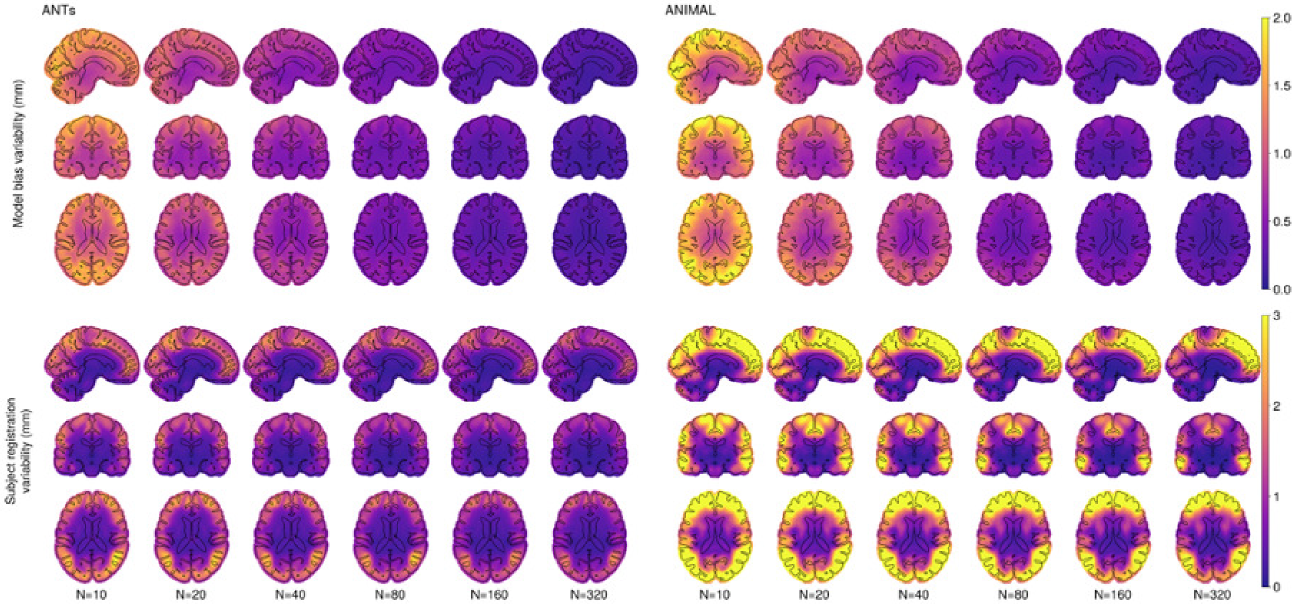
ANTs
ANIMAL
Model bias variability (mm)
Subject registration variability (mm)
2.0
1.5
1.0
0.5
0.0
3
2
1
0
N=10
N=20
N=40
N=80
N=160
N=320

Figure 2.1. Model bias variability (top) and individual subject registration variability (bottom) for ANTs (left) and ANIMAL (right) methods for N= 10, 20, 40, 80 160 and 320 subjects.

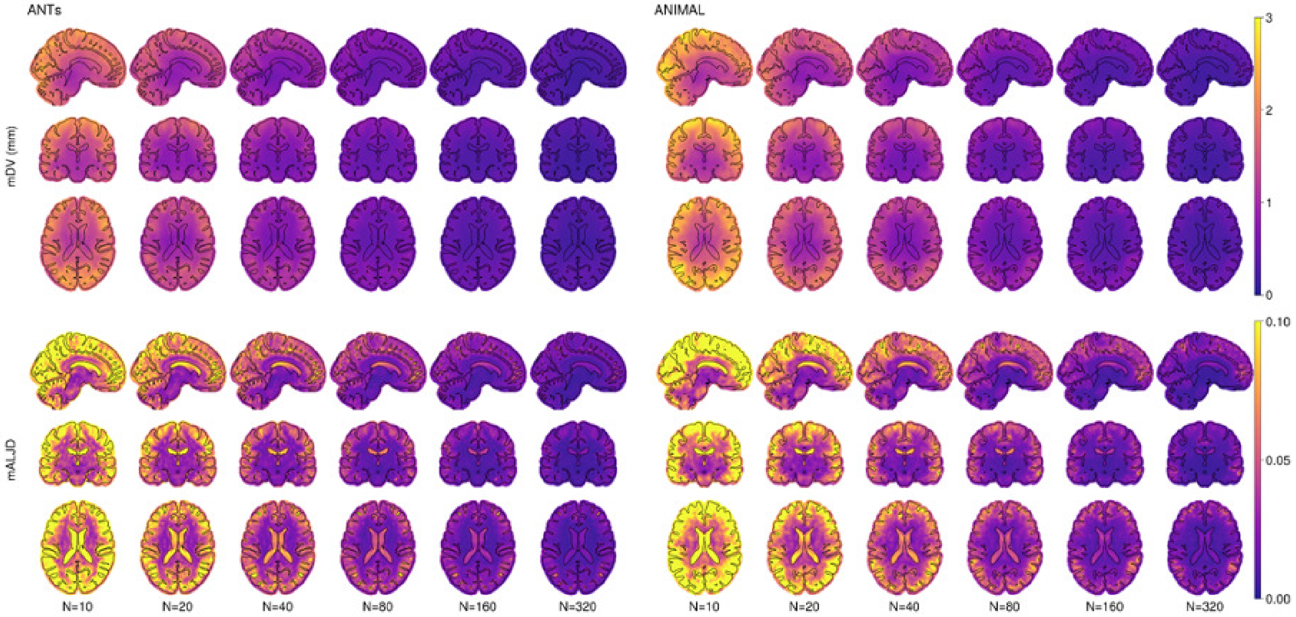

Figure 2.2. Mean deformation value (mDV, top) and mean absolute logarithmically transformed Jacobian determinant (mALJD, bottom) for ANTs (left) and ANIMAL (right) methods for N= 10, 20, 40, 80 160 and 320 subjects.

template. This may give an advantage to the ANTs strategy where non-linear registration of a subject benefits from the sharper edges in the unbiased average template target at each iterative step. Fourth, SynthSeg was used to segment regions in each subject's MRI. Any errors in segmentation will confound our inter- and intra-model overlap metrics used to estimate quality of registration. These confounds should affect ANTs and ANIMAL in a similar manner and not result in a bias between registration methods. In addition, inclusion of more subjects increases subject variability as the potential for contradictory anatomical information will reduce overlap and increase blurring. Fifth, our silver standard template is good, but not perfect. Since N=2000 is large, and we took care to include equal numbers of cognitively normal men and women, it should yield a good estimate of the population mean. However, these 2000 subjects were selected from the UK Biobank, which is already a subsample UK population only above 40 years of age. As seen from Table 2, 98% of the participants are white, and 84% are born in England. We expect our results to generalize to other healthy populations, but further experiments are needed to confirm this. We would also expect differ-

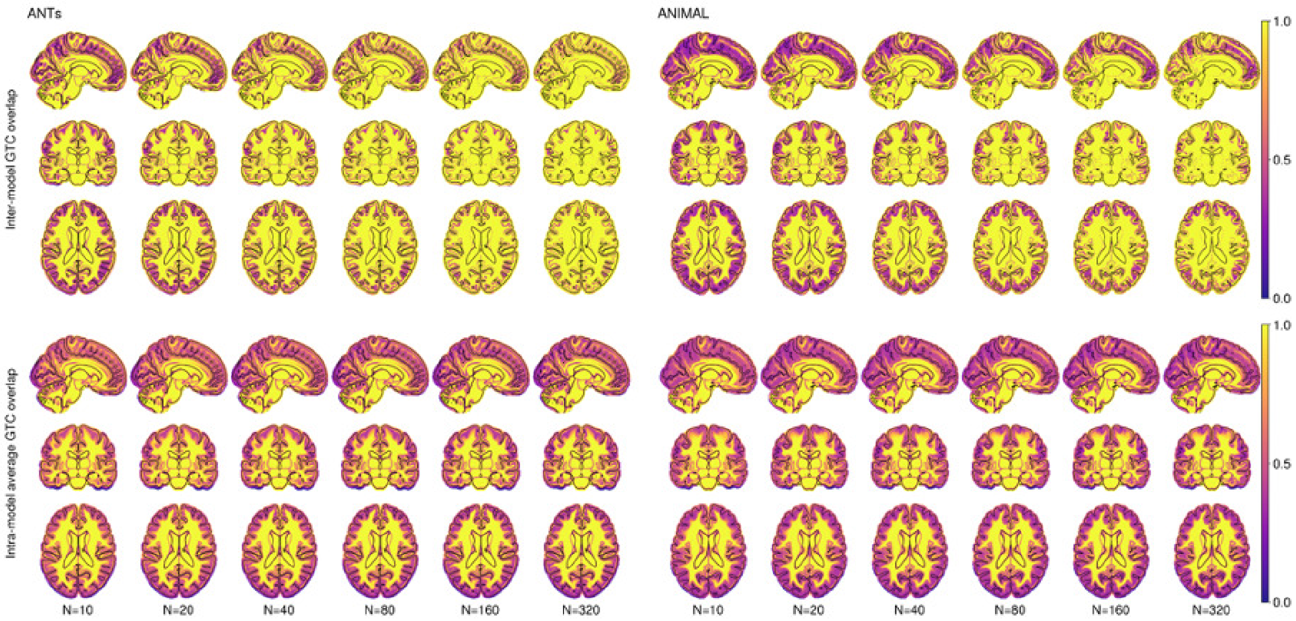

Figure 2.3. Inter-model overlap (top) and intra-model overlap (bottom) for ANTs (left) and ANIMAL (right) methods for N= 10, 20, 40, 80 160 and 320 subjects.

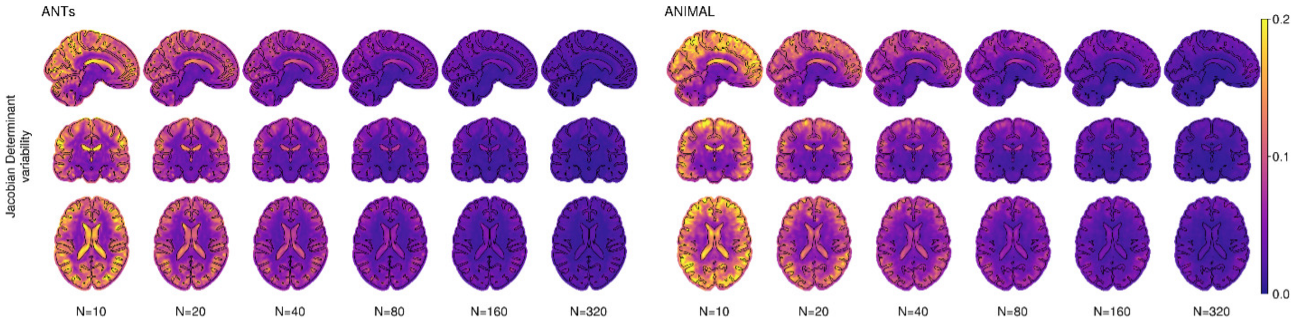

Figure 2.4. Jacobian determinant variability estimated for ANTs (left) and ANIMAL (right) methods for N= 10, 20, 40, 80 160 and 320 subjects.

ent numbers needed to achieve a stable population average when a diseased or more diverse population is being studied.

Neither ANTs nor ANIMAL can fully capture cortical variability—neither method can unfold the cortical anatomy of one subject to refold it to fit the gyral and sulcal pattern of another subject (or template). This suggests that surface-based methods may be better at precise cortical alignment to model the average cortex. While this might be possible for the primary and many secondary gyri and sulci, it is not clear that all secondary and tertiary gyri and sulci can be aligned because there is not a one-to-one homology across cortices for all subjects. In future work, it might be interesting to classify subjects into a particular cortical folding pattern for a specific sub-region of the cortex, and then average only subjects with the same folding pattern—but only for that specific region.

CONCLUSIONS

Our goal was to determine, within the population represented by the UK Biobank, what is the minimal number of subjects required to generate a stable average, where adding more subjects does not significantly improve the estimate of the average shape. In future work, it would be interesting to see if our results hold for sub-populations: for example, for men vs. women, for subjects with different neurodegenerative diseases such as Parkinson's or Alzheimer's disease or for adolescents with attention deficit hyperactivity disorder. One might hypothesize that for some disease-specific population templates, more subjects might be necessary due to added variability from disease, in addition to naturally occurring anatomical variability between subjects.

Our experiments show that 160 samples are sufficient to achieve a stable population average with both template building methods. However, if a smaller number of samples are available, ANTs achieves stable results with a smaller

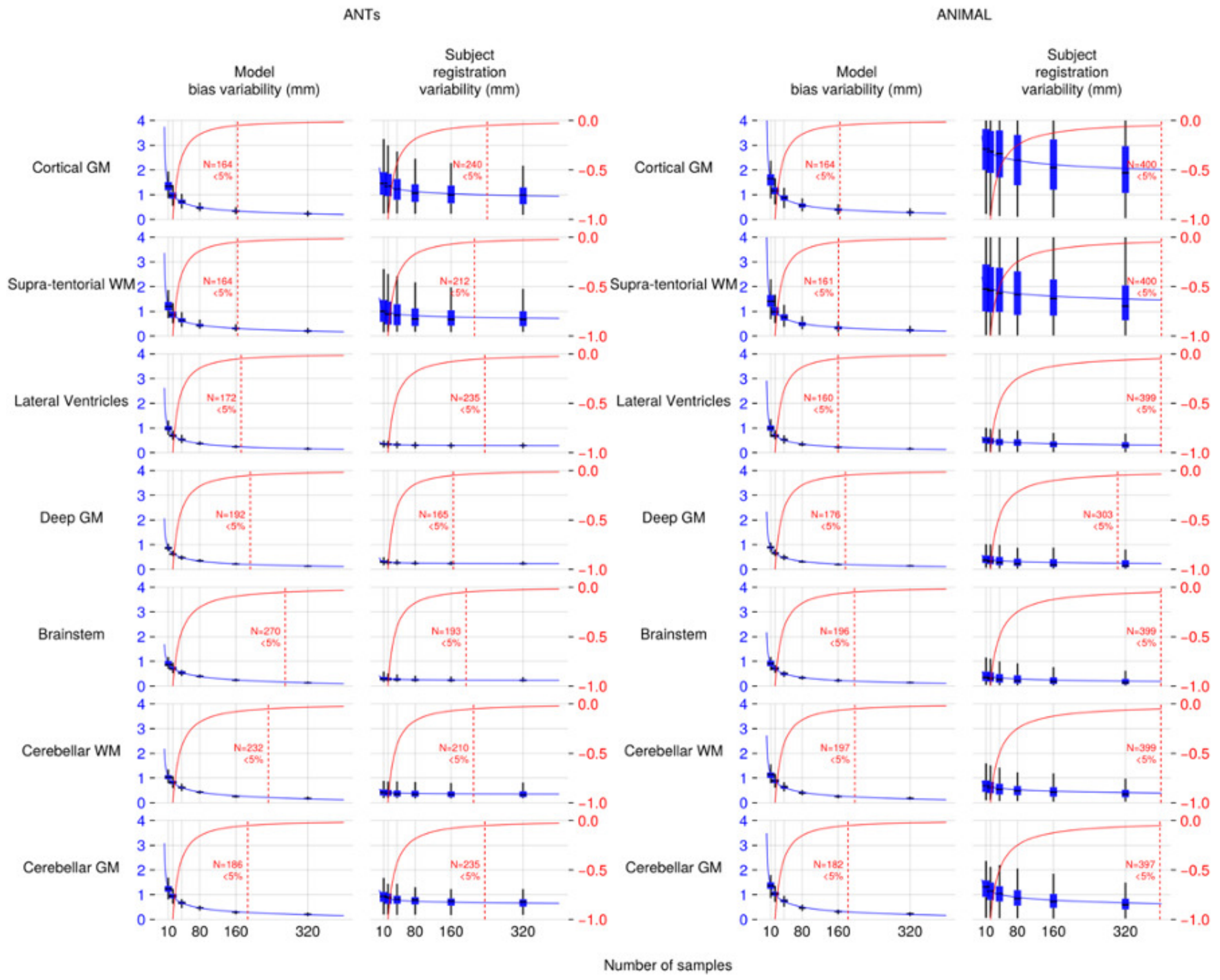

Figure 3.1. Model bias variability and individual subject registration variability for both unbiased template registration methods computed for different brain regions. The brain stem, deep grey matter nuclei, and ventricles show the lowest average position variability, and cortical grey matter the highest. Values for model bias variability are similar for ANTs and ANIMAL methods. Roughly half as many subjects are required for ANTs to achieve equivalent subject registration variability.

number of samples in most brain regions. Stability of the population average varies spatially, and neither method fully captures anatomical differences in the cortical folding patterns, indicating that surface-based or hybrid methods should be developed for the tasks where accurate cortical co-registration is needed. The silver standard average templates are available at: https://nist.mni.mcgill.ca/uk-biobank-average-2000/.

DATA AND CODE AVAILABILITY

Data used in the research are available from the UK Biobank: www.ukbiobank.ac.uk. The final versions of the generated templates are available at https://nist.mni.mcgill.ca/uk-biobank-average-2000/. Code used to generate average templates and to perform statistical analysis is available upon reasonable request.

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource under Application Number 45551.

This research used the NeuroHub infrastructure and was undertaken thanks in part to funding from the Canada First Research Excellence Fund, awarded through the Healthy Brains, Healthy Lives initiative at McGill University.

This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada.

We acknowledge the help of Dr. Stanley Hum, with editing the manuscript.

FUNDING SOURCES

Authors received funding from Famille Louise & André Charron and Canadian Institutes of Health Research (MOP-111169).

CONFLICTS OF INTEREST

Authors have no conflicts of interest to disclose.

Submitted: October 08, 2024 CDT. Accepted: January 31, 2025 CDT. Published: March 10, 2025 CDT.

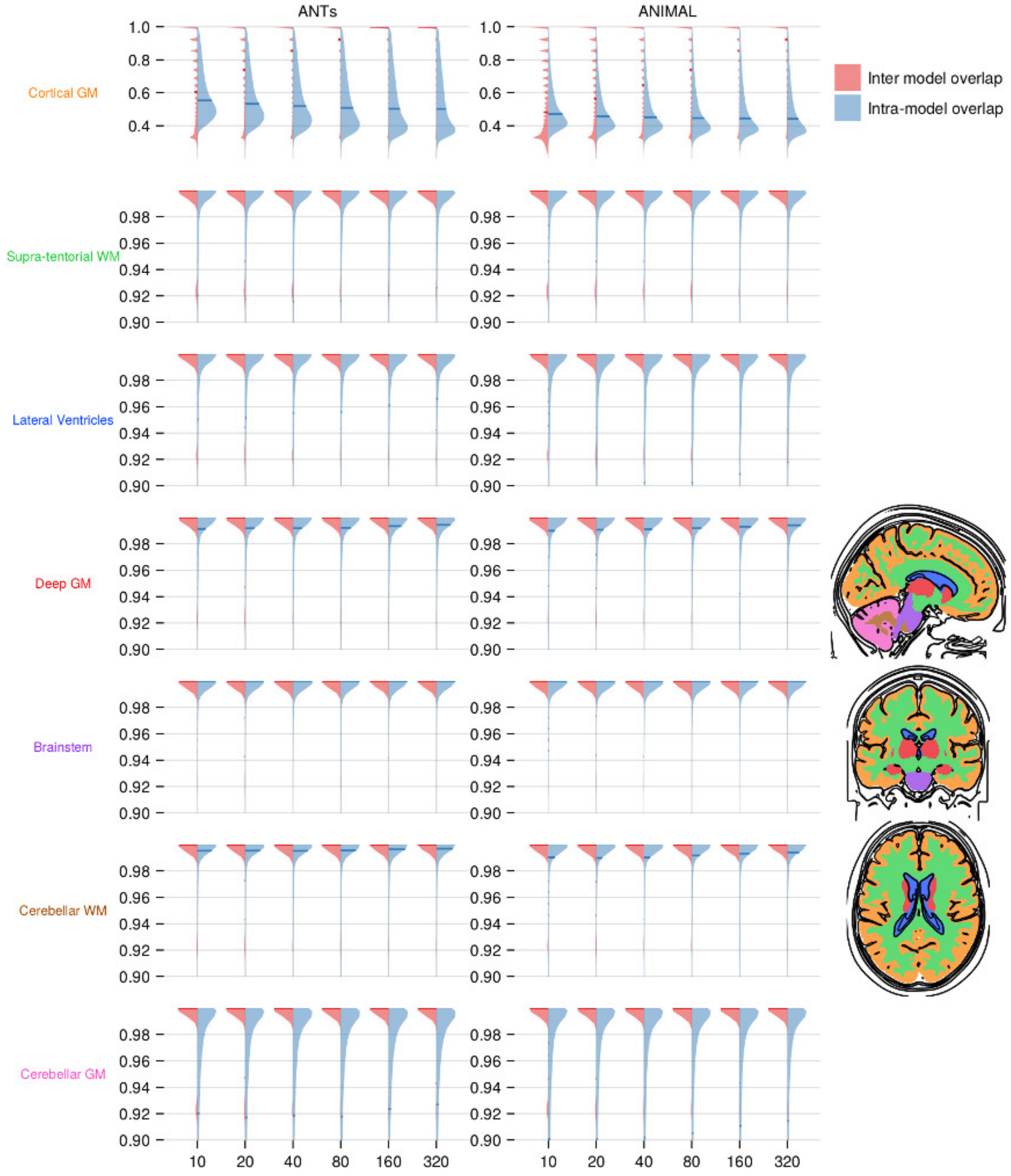

Figure 3.2. Inter-model overlap and intra-model overlap for both unbiased template registration methods computed for different brain regions. The brain stem, deep grey matter nuclei, and ventricles show the lowest average position variability, and cortical grey matter the highest.

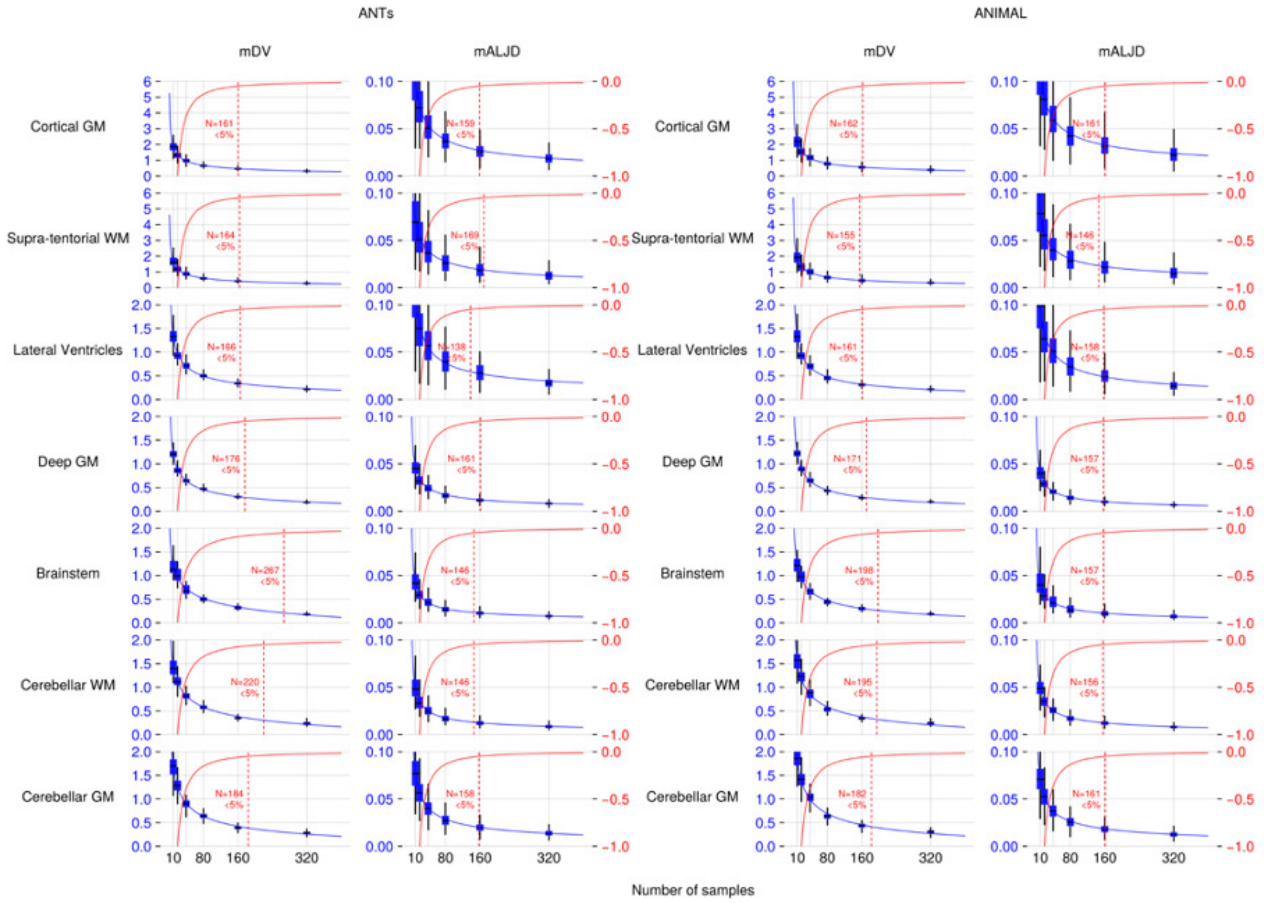

Figure 3.3. Mean deformation value (mDV) and mean absolute logarithmically transformed Jacobian determinant (mALJD) for ANTs (left) and ANIMAL (right) methods for N= 10, 20, 40, 80 160 and 320 subjects. These metrics show surprisingly similar results across all structures, except for the brainstem in ANTs, where the method converges with more subjects.

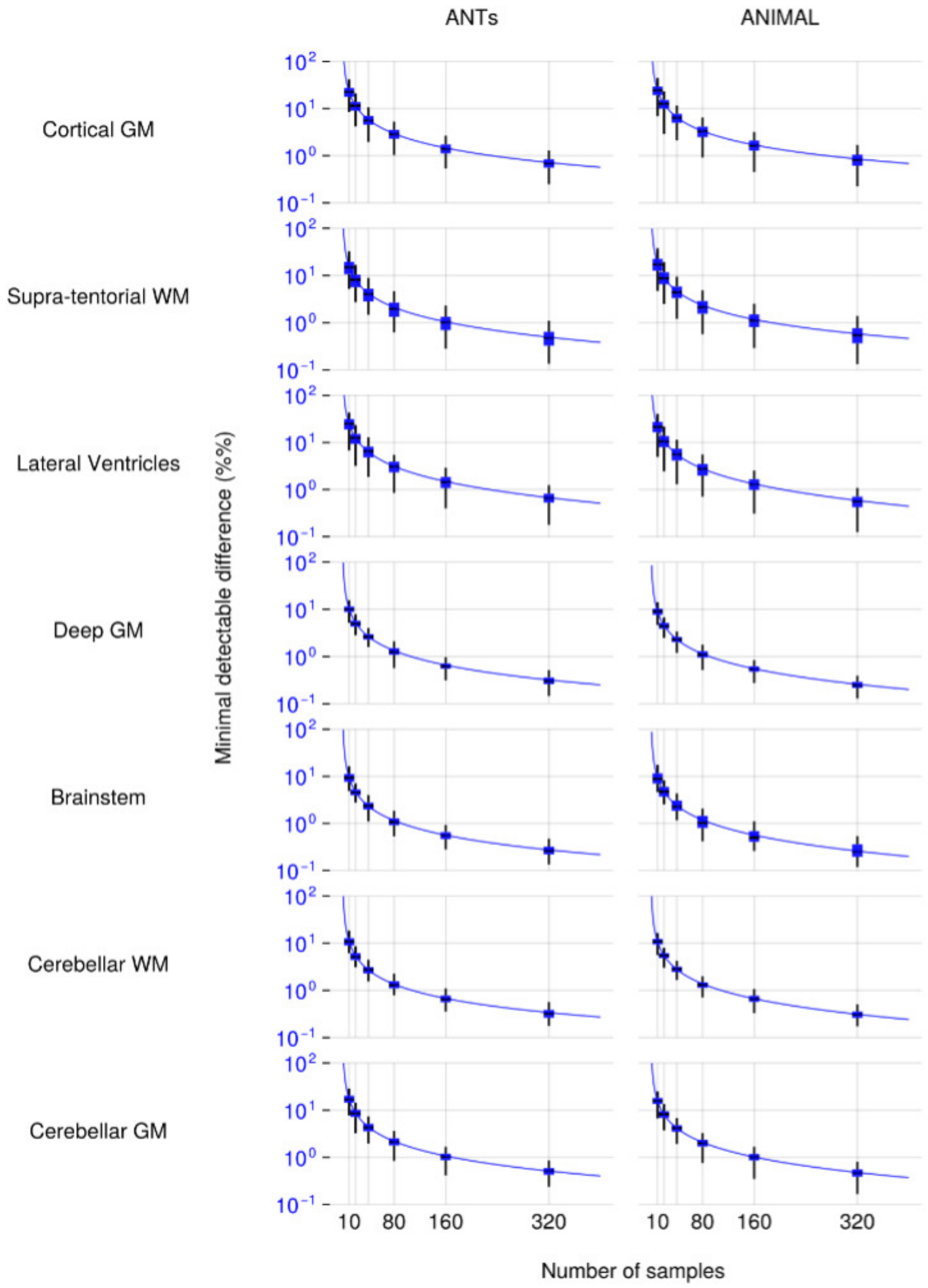

Figure 3.4. Estimation of the minimal detectable difference (in percent) for hypothetical deformation-based morphometry experiment shown on a log-scale for ANTs (left) and ANIMAL (right) methods for N= 10, 20, 40, 80 160 and 320 subjects. Here it is visible that ANTs allows detection of a given difference using a slightly smaller number of subjects.

REFERENCES

1. Evans AC, Janke AL, Collins DL, Baillet S. Brain templates and atlases. *NeuroImage*. 2012;62(2):911-922. doi:10.1016/j.neuroimage.2012.01.024

2. Fischl B. Automatically Parcellating the Human Cerebral Cortex. *Cereb Cortex*. 2004;14(1):11-22. doi:10.1093/cercor/bhg087

3. Friston KJ, ed. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. 1st ed. Elsevier/Academic Press; 2007. doi:10.1016/B978-012372560-8/50002-4

4. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. *NeuroImage*. 2012;62(2):782-790. doi:10.1016/j.neuroimage.2011.09.015

5. Mazziotta J, Toga A, Evans A, et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos Trans R Soc Lond B Biol Sci*. 2001;356(1412):1293-1322. doi:10.1098/rstb.2001.0915

6. Ciric R, Thompson WH, Lorenz R, et al. TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nat Methods*. 2022;19(12):1568-1571. doi:10.1038/s41592-022-01681-2

7. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*. 2011;54(1):313-327. doi:10.1016/j.neuroimage.2010.07.033

8. Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, Evans AC. Enhancement of MR Images Using Registration for Signal Averaging. *J Comput Assist Tomogr*. 1998;22(2):324-333. doi:10.1097/00004728-199803000-00032

9. Fonov V, Evans A, McKinstry R, Almli C, Collins D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*. 2009;47:S102. doi:10.1016/S1053-8119(09)70884-5

10. Avants B, Tustison N. ANTs/ANTsR Brain Templates. 2018. https://figshare.com/articles/dataset/ANTs_ANTsR_Brain_Templates/915436

11. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*. 2011;54(3):2033-2044. doi:10.1016/j.neuroimage.2010.09.025

12. Shi F, Yap PT, Wu G, et al. Infant Brain Atlases from Neonates to 1- and 2-Year-Olds. Okazawa H, ed. *PLoS ONE*. 2011;6(4):e18746. doi:10.1371/journal.pone.0018746

13. Wu G, Wang Q, Jia H, Shen D. Feature-based groupwise registration by hierarchical anatomical correspondence detection. *Hum Brain Mapp*. 2012;33(2):253-271. doi:10.1002/hbm.21209

14. Satterthwaite TD, Connolly JJ, Ruparel K, et al. The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage*. 2016;124:1115-1119. doi:10.1016/j.neuroimage.2015.03.056

15. Ciric R, Wolf DH, Power JD, et al. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*. 2017;154:174-187. doi:10.1016/j.neuroimage.2017.03.020

16. Avants B, Tustison N, Song G, Cook P, Klein A, Gee J. The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*. 2010;49(3):2457-2466. doi:10.1016/j.neuroimage.2009.09.062

17. Crum WR, Camara O, Hill DLG. Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. *IEEE Trans Med Imaging*. 2006;25(11):1451-1461. doi:10.1109/TMI.2006.880587

18. Ashburner J, Friston KJ. Computing average shaped tissue probability templates. *NeuroImage*. 2009;45(2):333-341. doi:10.1016/j.neuroimage.2008.12.008

19. Yang G, Zhou S, Bozek J, et al. Sample sizes and population differences in brain template construction. *NeuroImage*. 2020;206:116318. doi:10.1016/j.neuroimage.2019.116318

20. Van Essen DC, Ugurbil K, Auerbach E, et al. The Human Connectome Project: A data acquisition perspective. *NeuroImage*. 2012;62(4):2222-2231. doi:10.1016/j.neuroimage.2012.02.018

21. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

22. Fonov VS, Dadar M, The PREVENT-AD Research Group, ADNI, Collins DL. DARQ: Deep learning of quality control for stereotaxic registration of human brain MRI. Published online August 18, 2021. doi:10.1101/2021.08.16.456514

23. Dadar M, Fonov VS, Collins DL. A comparison of publicly available linear MRI stereotaxic registration techniques. *NeuroImage*. 2018;174:191-200. doi:10.1016/j.neuroimage.2018.03.025

24. Collins DL, Evans AC. Animal: Validation and Applications of Nonlinear Registration-Based Segmentation. *Int J Pattern Recognit Artif Intell*. 1997;11(08):1271-1294. doi:10.1142/S0218001497000597

25. Avants B, Epstein C, Grossman M, Gee J. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*. 2008;12(1):26-41. doi:10.1016/j.media.2007.06.004

26. Billot B, Greve DN, Puonti O, et al. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Med Image Anal*. 2023;86:102789. doi:10.1016/j.media.2023.102789

27. van Belle G. *Statistical Rules of Thumb*. 2nd ed. John Wiley & Sons, Inc.; 2011. doi:10.1002/9780470377963