The Effect of Training Dataset Size on the Quality of MRI Site Harmonization

Heuvel, Liam van den

The Effect of Training Dataset Size on the Quality of MRI Site Harmonization

Master's Thesis

Liam van den Heuvel

Master's Thesis

Methodology and Statistics Unit, Institute of Psychology,

Faculty of Social and Behavioral Sciences, Leiden University

Date: March 2023

Student number: s1800345

Supervisor: Dr. Frank de Vos (internal)

Acknowledgements

I would like to express my gratitude to Dr. Frank de Vos for his indispensable guidance throughout the process of writing this thesis and his valuable feedback. I would also like to give thanks to Lude Rozema for generously providing me background information on the topic of this thesis. Lastly, I would like to thank my girlfriend, my family and my friends for their support and motivation during the process of writing this thesis.

Abstract

Most neuroimaging studies are affected by small sample sizes and poor reproducibility of research findings. Therefore, aggregating data from multiple research centres is crucial to the development of the neuroimaging field. For this reason, MRI site harmonization is essential, as it allows for comparison and joint analysis of MRI data from multiple studies. MRI site harmonization aims to remove inter-site variability, while maintaining variance of interest. However, neuroimaging studies generally have low numbers of subjects to estimate the harmonization model. This paper examines the effect of dataset size on the quality of MRI site harmonization, and whether this effect is dependent on age differences between sites and the size of site differences. In order to evaluate the quality of MRI site harmonization we calculated the extent to which the correlation between GMD and age was recovered. To answer our research questions, we studied the performance of MRI site harmonization using a variety of training dataset sizes in an empirical study. Our empirical study shows no clear effect of the size of the training dataset. In addition, we studied the performance of MRI site harmonization in a simulation study, where we varied the number of subjects in the training dataset, the age differences between the centres, and the size of the centre effects. Our simulation study shows that the effect of training dataset size is minimal. The effect is only present when sites differ largely in mean age and when site effects are small. Thus, in all other conditions, inter-site variability is successfully removed, while variance of interest is preserved. This leads us to the conclusion that the limited effect of training dataset size suggests that prospects for the quality of harmonization in multi-centre studies with small datasets are promising.

Table of Contents

Introduction 4
Research questions 5
Method 6
Empirical Application 6
Participants 6
Missing data 6
MRI Acquisition 7
MRI Analysis 7
Centre effects 7
Simulation study 10
Procedure 12
Evaluation 13
Results 14
Empirical Application 14
Simulation study 15
Research question 1 15
Research question 2 17
Research question 3 21
Discussion 24
References 27

Introduction

Recently, concerns have been raised in the field of neuroimaging as a result of small sample sizes and poor reproducibility of research findings, among other things (Zhu et al., 2019). Since many studies that make use of MRI data are bound to small sample sizes, aggregation of MRI data from multiple studies is crucial to the development of the neuroimaging field. In the field of neuroimaging, MRI data is subject to significant inter-site variability. Variability introduced by processing data at different sites is referred to as "batch effects" (Chen et al., 2011). These batch effects evidently complicate comparison of data across scanning sites (Pinto et al., 2020). In recent years, applied researchers have become increasingly interested in reducing inter-site variability. A standard procedure for reducing the inter-site variability is the harmonization of data, which removes variability due to scanner site and maintains variability due to biological and demographical factors. Harmonization therefore allows for comparison and joint analysis of MRI data from multiple studies. Several studies have covered the application of different harmonization methods on MRI data; see Chen et al., 2011; Nan et al., 2022 for a review of multiple harmonization methods on several imaging modalities, including MRI. It has been demonstrated that the ComBat harmonization technique is a promising harmonization method. ComBat adjusts the data by removing site effects using linear regression, while adjusting for known covariates. In order to make the variance similar across sites, ComBat add site-specific scaling factors. ComBat also uses empirical bayes to improve the estimation of site parameters for small sample sizes. Although the ComBat harmonization technique is promising, datasets from most neuroimaging studies are generally small, and possibly too small to adequately estimate the harmonization model. To date, little attention has been devoted to the quality of MRI site harmonization for small datasets. The aim of this thesis is to evaluate the effect of the training dataset size on the quality of MRI site harmonization. To this end, a data simulation study will be performed to examine the effect of the training dataset size on MRI site harmonization. A simulation study is used to test the behaviour of the ComBat model under controlled conditions. Additionally, I will accompany the data simulation study with an empirical application.

Research questions

In this thesis, I will examine whether the training dataset size affects the quality of MRI site harmonization in an empirical dataset. Next, I will examine the same matter in a simulation study. I hypothesize that the larger the size of the training dataset, the better the performance of MRI site harmonization, since the use of more training data adds information to the harmonization model. In addition, I will examine whether the effect of training dataset size is

dependent on between-site differences in mean age. By increasing the demographical variance in the data, it will be more difficult for ComBat to differentiate between demographic and site-related variance. This increases the complexity of the harmonization model and might therefore require more data to accurately train the model. Also, I will examine whether the effect of training dataset size is dependent on the size of the site effects in the dataset. I hypothesize that the effect of training dataset size is larger when site-related variance is large, as the increase of inter-site variability will increase the complexity of the model. As a consequence of this, the model will require more data.

Method

Empirical Application

Participants

A multi-site dataset from the Leiden Alzheimer Research Nederland (LeARN) project (Handels et al., 2012; Jansen et al.,2017) was used in this study, which was collected at four memory clinics in the Netherlands; Leiden, Maastricht, Nijmegen and Amsterdam. The LeARN dataset consisted of 61 possible or probable Alzheimer's Disease (AD) patients, 61 Mild Cognitive Impairment (MCI) patients and 67 Significant Memory Concern (SMC) patients, resulting in a total of 189 subjects.

Table 1 gives an overview of the demographic characteristics of the sample. This dataset was chosen because it was readily available and contains MRI data from multiple research centres. With this empirical dataset, I examined the effect of training dataset size on the quality of MRI site harmonization. In addition, I used characteristics of this dataset as a starting point to simulate random samples in the simulation study.

Missing data

The percentage of missing data across all variables in the empirical dataset varied between 0 and 10%. Therefore, a total of 32 of 10206 data points were incomplete. Important to note is that no MRI variables contained missing data, and missing data was only found on covariates. It is essential to the ComBat harmonization technique that there is no missing data. Therefore, missing data was handled by imputation of the median. Acuña & Rodriguez (2004) showed that missing data should be handled by imputation of the median when the data in the variables containing missing data is skewed. The skewness found on the GDS and MMSE variables could likely be attributed to floor- and ceiling effects respectively.

MRI Acquisition

Subjects underwent on-site structural MRI scans, calculating grey matter density (GMD) in 48 locations in the brain, representing the 48 cortical regions of the probabilistic Harvard-Oxford cortical atlas (Smith et al., 2004). Scanners were site-specific; a Philips Achieva 3T scanner was used at the Leiden University Medical Centre and the Maastricht University Medical Centre, a Siemens TrioTim 3T was used at the Nijmegen University Medical Centre, and a GE Signa HDxt 3T scanner at the VU University Medical Centre in Amsterdam.

MRI Analysis

Grey matter density represents the percentage of grey matter, which is quantified by a number between 0 and 1. This is achieved by soft segmentation of brain voxels into grey matter, white matter or cerebral spinal fluid (Gennatas et al., 2017). The voxel grey matter density values were then averaged within each of the 48 cortical regions. See de Vos et al., 2020 for further information on the MRI data acquisition and data processing of this dataset.

Centre effects

Due to the subjects being scanned at different sites, inter-site variability was likely to be present in the data. In other words, means and standard deviations of the GMD variables were subject to a centre effect introduced by the different scanners. The harmonization process tends to remove these centre effects. To visualize this, Figure 1 shows the GMD values of all subjects both before and after harmonization. The x-axis represents the 189 subjects present in the empirical dataset. The y-axis represents the 48 cortical regions of the probabilistic Harvard-Oxford cortical atlas. Comparison of the two plots in Figure 1 showed that site effects were present in the data. The scan site correction process was not biased towards one direction, as the correction was different for every variable, and site specific. To assess the size of the centre effect that is present in the empirical data, two sets of 48 one-way ANOVAs were conducted to calculate the effect of centre on GMD in all 48 cortical brain regions. We conducted two sets of 48 ANOVAs to differentiate between variability due to centre effects and variance of interest, as in one set of ANOVAs several covariates were included. These ANOVAs revealed that the mean effect size of centre, excluding covariates, before harmonization was $\eta^2 = .26$. However, the mean effect size of centre including covariates (i.e., age, CDR, years of education, GDS, MMSE, diagnosis and sex) was $\eta^2 = .31$. These results seem odd, as we expect a decrease in effect size since the addition of covariates explains extra error variance. However, Table 1

shows us that the four centres differ only slightly on most covariates. The distribution of the centre effect sizes both with and without the addition of covariates, and both before and after harmonization are displayed in Figure 2.

Table 1

Demographic Characteristics of the LeARN Dataset

Characteristic	Scanning site			
	Leiden	Maastricht	Nijmegen	Amsterdam
N	40	68	43	38
Age	70.9 ± 9.0	66.6 ± 11.6	71.6 ± 9.0	65.0 ± 7.5
CDR	.58 ± .27	.52 ± .16	.49 ± .37	.62 ± .32
Years of education	11.1 ± 3.7	10.4 ± 3.2	11.1 ± 3.7	11.6 ± 3.4
GDS	3.8 ± 3.1	3.26 ± 2.5	2.79 ± 1.8	3.44 ± 2.8
MMSE	26.4 ± 2.5	27.5 ± 2.6	25.7 ± 2.8	25.4 ± 3.0
Diagnosis				
AD	15	13	14	19
MCI	13	24	13	11
SMC	12	31	16	8
Sex				
Male	20	42	25	30
Female	20	26	18	8

Note. CDR = clinical dementia rating, GDS = geriatric depression scale, MMSE = mini-mental state examination, AD = Alzheimer's disease, MCI = Mild cognitive impairment, SMC = Significant memory concern. Age, CDR, Years of education, GDS and MMSE are presented as mean ± standard deviation. Diagnosis and Sex are presented as frequencies.

Figure 1

Result of the harmonization process. The 48 scanned brain regions are represented on the y-axis. The 189 subjects are represented on the x-axis, sorted by site.

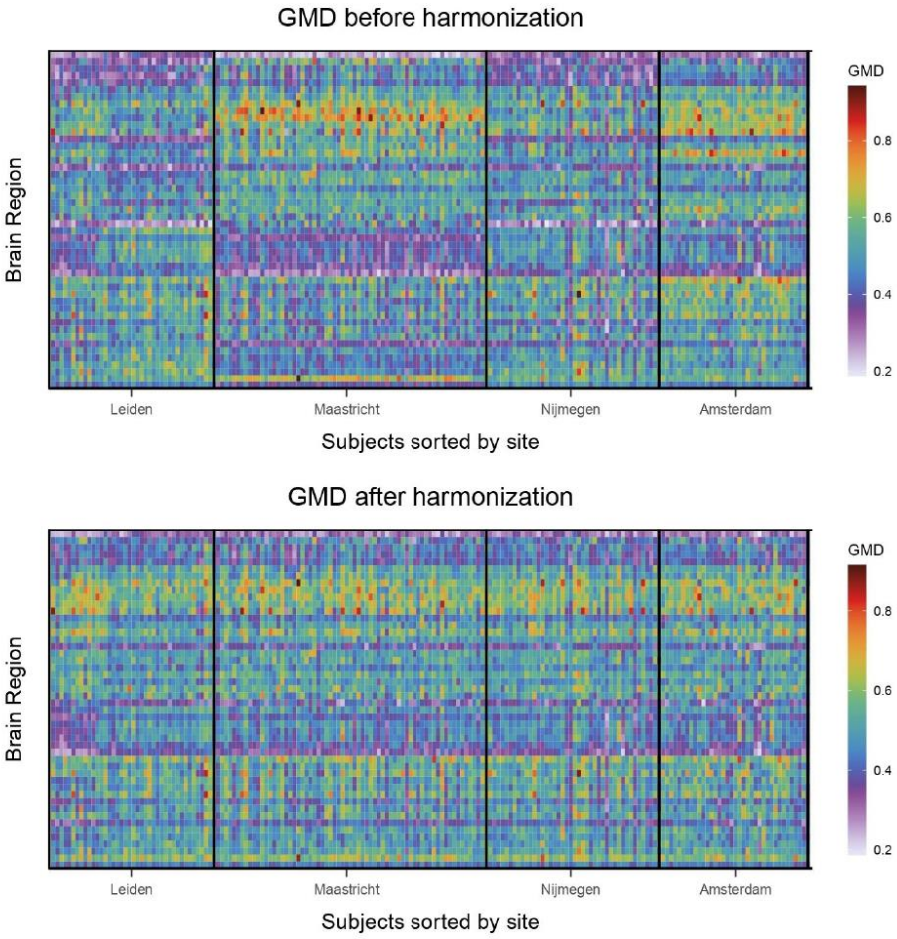
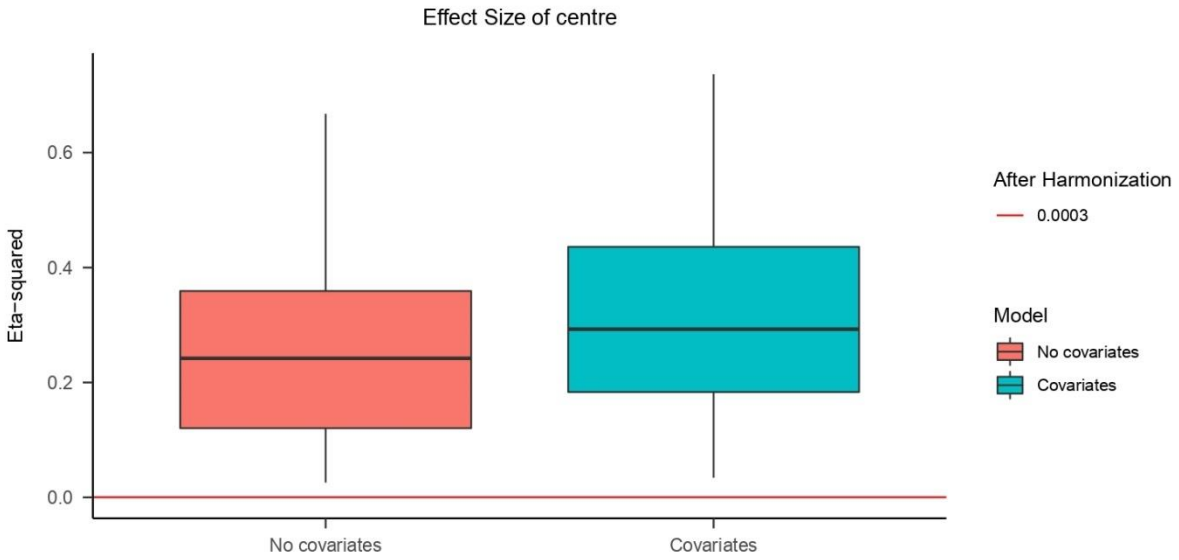


Figure 2

Distribution of the centre effect sizes. The boxplot on the left represents the effect sizes of 48 one-way ANOVAs of the effect of centre on GMD excluding covariates. The boxplot on the right represents the effect sizes of 48 one-way ANOVAs of the effect of centre on GMD including covariates. The y-axis represents the eta-squared effect size. The red line represents the mean centre effect sizes after harmonization.



Simulation study

Data was simulated, for each centre separately, for a total of 1000 repetitions, from a multivariate normal distribution: $X_{ijk} \sim N(\mu_{ij}, \Sigma_{ij})$, where μ_{ij} is the mean vector of the 35 GMD variables and an age variable of centre 1 up to 4 in the empirical dataset, and Σ_{ij} is the $n \times n$ positive definite covariance matrix of centre 1 up to 4 in the empirical dataset. Data was simulated using the faux package (v1.1.0; DeBruine., 2021). Thus, four centre-specific datasets were created, which were aggregated after every repetition to form one multi-centre dataset. In short, a dataset consisted of four centres, each containing 100 subjects, resulting in a total of 400 subjects per dataset. Each subject had 35 variables representing GMD in 35 brain regions and an age variable, which results in a total of 36 variables.

We wanted the variable properties to be plausible and realistic, hence the use of the parameters from the empirical dataset as population parameters. For each subject, 35 variables representing GMD in different brain regions along with an age variable were simulated. We simulated 35 GMD variables, as opposed to 48 variables in the empirical dataset, for the reason that the faux package required more subjects per centre in the empirical dataset than simulated variables to produce a dataset. Since the lowest number of subjects scanned at a specific centre

was 38 (at the VU University Medical Centre in Amsterdam), no more than 38 variables could be simulated. We decided to use 35 GMD variables as this is a round number. The simulation study was performed to answer three research questions, for which data-simulation parameters were altered accordingly.

Firstly, in order to answer our second research question, age differences between sites were given to the data. Specifically, the mean age value in the mean vectors of two of the four centres were altered to introduce an age difference of 10 or 20 years. The mean age value in the mean vectors of the remaining two centres was 60. Thus, two centres had a mean age of 60, and two centres had a mean age of either 70 or 80, causing a small and large age difference respectively. As mentioned above, GMD is negatively correlated with age. Therefore, it was of importance that GMD values were corrected for the increase of age. For every centre, we calculated the effect of age on each of the 35 simulated GMD variables. Next, we created a matrix of the age effect on all 35 GMD variables for all subjects. At the end, this matrix was added to the unaltered simulated data to account for the increase in age, resulting in an age-corrected dataset.

Secondly, to answer our third research question, site effects were manipulated by multiplying, for every centre separately, the mean vector of GMD in 35 brain regions with a factor. This factor was different for every condition, depending on the intended size of the site effect. As a consequence of this, the between centre variance decreased when this factor was smaller than 1, and increased when the factor was larger than 1. We chose multiplication factors that resulted in datasets with a desired centre effect sizes (i.e., eta-squared) of .01, .06, .14, .26, corresponding to a small, medium and large effect defined by Cohen (1988). In addition, we simulated data with an eta-squared of .26, as this was the mean effect size found in the empirical application excluding covariates. We will refer to this effect size as ‘huge’ in the following sections.

An age variable was simulated to aid in the evaluation of harmonization success. This evaluation is based on previous research (Ramanoël et al., 2018), which found that an increase in age is associated with a decrease in GMD. Therefore, if the correlation between GMD and age increased after harmonization, we concluded that the harmonization removed, at least, some of the centre effect. Since it is unclear whether remaining site differences are a result of an insufficient harmonization process, or result from demographical and/or biological variance, objectively estimating harmonization success is rather difficult. Therefore, the estimation of an intended effect (i.e., the correlation between GMD and age) is perceived as the golden standard for the determination of harmonization success. As a reference, Table 2 shows the mean,

standard deviation and range of the simulated mean correlation between GMD and age for all 1000 simulated datasets, which gives a general idea of the intended effect.

Table 2

Descriptive statistics of the mean correlation between GMD and age for all simulated datasets

	M	*SD*	*Min*	*Max*
Centre 1	*-.38*	*.04*	*-.50*	*-.20*
Centre 2	*-.46*	*.04*	*-.55*	*-.30*
Centre 3	*-.47*	*.05*	*-.61*	*-.30*
Centre 4	*-.09*	*.04*	*-.24*	*.06*

Procedure

After data simulation, these data were subjected to a MRI site harmonization method called ComBat (Johnson et al., 2007). The ComBat harmonization technique has been found to successfully remove unwanted site-related variability, while preserving biological and demographic variability in the data (Fortin et al., 2017). ComBat builds on other harmonization procedures that remove site effects using linear regression, while adjusting for known covariates. Additionally, ComBat adds site-specific scaling factors in order to make the variances similar across sites. Furthermore, it uses empirical bayes to improve the estimation of the site parameters for small sample sizes (Fortin et al., 2018). In spite of the fact that ComBat was originally developed to mitigate non-biological variability in gene expression microarray data (Johnson et al., 2007), recently ComBat has been further developed to mitigate batch effects in MRI data (Fortin et al., 2017). In this study a modified version of the 'ComBat' function from the 'sva' package in R was used. In this package, the ComBat function was adapted in such a way that a function for fitting and applying the harmonization model could be applied separately (Radua et al., 2020). This enabled us to only use a subset of the subjects (size of training data) to fit the harmonization model, and apply it to all of the subjects using cross validation.

In the empirical application, the number of subjects in the training dataset had six conditions: 20%, 33%, 50%, 67%, 80% and 100% of the subjects per centre. Afterwards, the remaining subjects were used to validate the harmonization model using cross validation. The

model consisted of 48 GMD variables, a centre variable, and six covariates (i.e., age, CDR, years of education, GDS, MMSE and sex). We used a cross validation approach to accomplish that all subjects were used for testing the harmonization model.

The simulation study examined the effect of three factors. Firstly, the main effect of training dataset size on MRI site harmonization quality was examined. The number of subjects in the training dataset had eight conditions; 5, 10, 20, 33, 50, 67, 80 and 100 (all) subjects per centre. Subsequently, the remaining subjects were used to validate the harmonization model using a cross validation approach. Secondly, we examined whether the effect of training dataset size on MRI site harmonization quality was dependent on between-site differences in mean age. We created three conditions; no age difference, a ten-year age difference and a 20-year age difference. Thirdly, we examined whether the main effect of training dataset size on MRI site harmonization quality was dependent on the sizes of site effects within the data. We created four conditions; small site differences, medium site differences, large site differences and huge site differences. We used a harmonization model consisting of 35 GMD variables, a centre variable and an age variable for all conditions.

Evaluation

As mentioned above, an age variable was simulated along with 35 GMD variables. We calculated the correlation between GMD and age before harmonization. MRI site harmonization success was then evaluated by the extent to which the correlation between GMD and age was recovered after harmonization.

In the empirical application, we examined the effect of training dataset size on MRI site harmonization quality in a within-subjects ANOVA design using training dataset size as a within subject factor and pearson correlation values between age and the GMD variables as the outcome variable.

In the simulation study, the main effect of training dataset size on MRI site harmonization quality was examined in a within-subjects ANOVA design. Moreover, we examined whether the effect of training dataset size on MRI site harmonization quality was dependent on between-site differences in mean age in a mixed design (8 x 3) ANOVA. In this ANOVA the size of between-site age differences functioned as a between-subjects variable, the training dataset's size as a within-subjects variable, and the mean value of the pearson correlation between GMD and age of the 35 GMD variables for all repetitions as the outcome variable. To answer this research question, we tested the interaction effect between training dataset size and age difference. Furthermore, we examined whether the main effect of training

dataset size on MRI site harmonization quality was dependent on the sizes of site effects within the data in a mixed design (8 x 4) ANOVA. In this ANOVA the size of the site-effects functioned as a between-subjects variable, the training dataset's size as a within-subjects variable, and the mean value of the pearson correlation between GMD and age of the 35 GMD variables for all repetitions as the outcome variable. To answer this research question, we tested the interaction effect between training dataset size and the size of site differences. All effects were examined with an ANOVA using the rstatix package (v0.7.1; Kassambara., 2020).

To assess the size of these effects we used generalized eta-squared (η^2_G) as an effect size. Bakeman (2005) encouraged the use of η^2_G, due to its comparability across studies and its appliance to both within- and between-subjects designs. Furthermore, reporting of η^2_G is recommended by Lakens (2013), since η^2_G excludes variation from other factors, such as the inclusion of covariates, which makes it possible to compare the effect size with a design in which these factors were not manipulated. On the contrary, η^2_G includes variance resulting from individual differences, which makes η^2_G comparable with between-subjects designs where variance resulting from individual differences cannot be controlled for.

All source code of this thesis is available at https://github.com/liamvandenheuvel/Master-s-Thesis-Liam-van-den-Heuvel.

Results

Empirical Application

In this empirical study we examined the effect of training dataset size on the quality of MRI site harmonization. We calculated the correlations between GMD and age for 48 brain regions to evaluate the quality of MRI site harmonization. For each training dataset size we calculated the mean correlation over all 48 correlation values. We found a mean correlation of r = -.371 for the pre-harmonization condition. We also found that the mean correlation increased for all conditions after harmonization, however this effect was not significant (F (1, 47) = 3.59, p = .064, η^2 = .007). The strongest mean correlation after harmonization was found when 33% of the subjects per centre were used to train the data (r = -.434). The smallest mean correlation after harmonization was found when 80 subjects per centre were used to train the data (r = -.387). In addition, a repeated measures ANOVA was performed to compare the effect of training dataset size on the correlation between GMD and age, excluding the pre-harmonization condition. The repeated measures ANOVA, where the six training dataset sizes functioned as a within subjects factor, and the 48 brain regions as cases,

revealed that there was a difference in the correlation between GMD and age between at least two conditions (F (5, 235) = 11.04, p < .001, η^2 = .012). Besides there being no effect of harmonization on the correlation between GMD and age, no clear pattern can be found in the relation between the correlation value between GMD and age and the size of the training dataset. Figure 3 shows the correlations between GMD and age for all training dataset size conditions in the empirical study.

Figure 3

Correlation between GMD and age for different training dataset sizes, as a measure of harmonization success. The 'Pre' condition represents unharmonized data. The x-axis represents the percentage of subjects of the total sample in the training dataset. The y-axis represents the correlation between GMD and age. These boxplots represent 48 correlations per condition, corresponding to GMD in the 48 cortical brain regions in the Harvard-Oxford cortical atlas.

Simulation study

Research question 1

This research question focused on the effect of training dataset size on the quality of MRI site harmonization. We assessed this by using different sizes of training data, ranging from five subjects (i.e., 5%) per centre to 100 subjects (i.e., 100%) per centre to train the ComBat harmonization model. Figure 4 shows the results. The y-axis represents the average correlation between GMD and age over all 35 brain regions and for all 1000 simulation repetitions. The x-axis represents all training dataset size conditions. I hypothesized that a

larger training dataset resulted in a stronger correlation after harmonization, as you have more data at your disposal for harmonization. As can be seen, the correlation between GMD and age was approximately the same as training dataset size decreased. The strongest mean correlation (r = -.309) was found when 5 subjects per centre were used to train the data. The smallest mean correlation (r = -.301) was found when 67 subjects per centre were used to train the data. A repeated measures ANOVA was performed to compare the effect of training dataset size on GMD, excluding the unharmonized data. This repeated measures ANOVA revealed that there was no difference in GMD between at least two conditions (F (7, 6993) = 1.295, p = .248, η^2_g = .001). Table 3 shows the means and standard deviations of the correlations between GMD and age for all training dataset sizes and Figure 4 shows the same results in boxplots. Apparently, harmonization with small datasets is as successful as harmonization with larger datasets. Moreover, all training dataset size conditions outperformed the non-harmonized sample (F (1, 999) = 26616.4, p < .001, η^2_g = .872). Although means hardly differ between the conditions, there seems to be an increase in variance in conditions where the training dataset size was smaller than 20 subjects per centre, which indicates that the quality of MRI site harmonization becomes precarious when there is not enough data to train the harmonization model. Additionally, Figure 4 could suggest that harmonization caused a decrease in mean correlation for some repetitions, since some of the boxplots of the harmonized results overlap with the boxplot of the 'pre' condition. Therefore, we calculated the percentage of increased correlations between GMD and age for all conditions compared to the situation without harmonization to rule out this explanation. We found that, for every repetition, the correlation between GMD and age increases after harmonization, meaning that harmonization was beneficial in all instances, even if the training dataset size is small.

Table 3

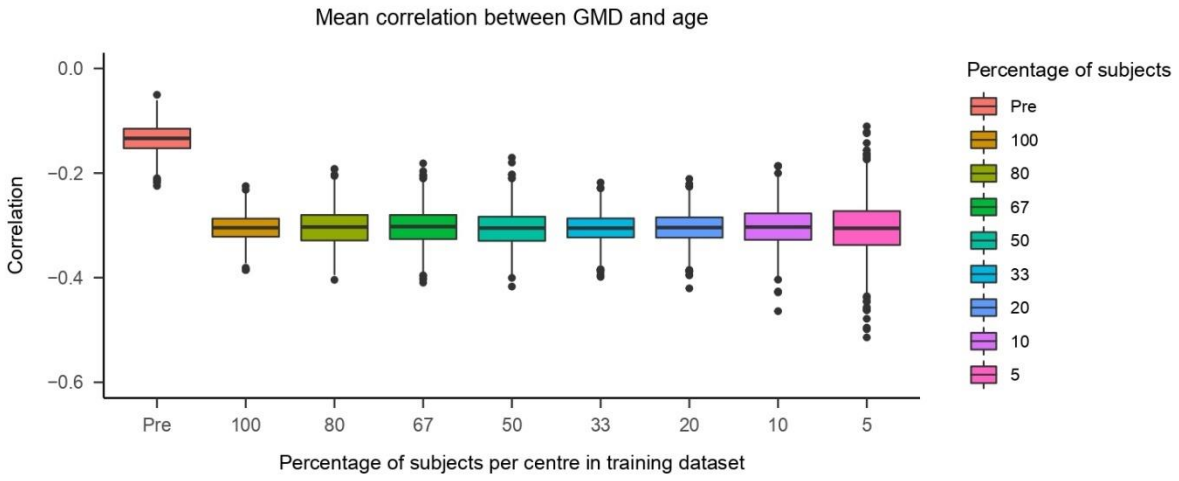
Descriptive statistics of the correlation between GMD and age for all training dataset sizes

	Subjects per centre in training dataset								
	Pre	*100*	*80*	*67*	*50*	*33*	*20*	*10*	*5*
Mean	-.134	-.301	-.301	-.302	-.302	-.302	-.302	-.304	-.309
SD	.028	.028	.030	.031	.032	.027	.029	.035	.046

Note. The "Pre" column represents descriptive statistics for the unharmonized data.

Figure 4

Mean correlation over all repetitions between GMD and age for different training dataset sizes. 'Pre' condition represents unharmonized data with 100 subjects per centre. The x-axis represents the number of subjects per centre in the training dataset. The y-axis represents the correlation between GMD and age.



Research question 2

To assess whether the effect of training dataset size on MRI site harmonization quality was dependent on age differences between sites, we created three settings that differed in their difference between the mean age of the centres. In the first setting, the mean age of subjects was equal between centres. In the second setting, we created a small mean age difference (10 years) between the centres, and in the third setting we created a large mean age difference (20 years) between the centres. Figure 5 shows the correlation between GMD and age for all

training dataset sizes split out over the three settings. Table 4 shows the means and standard deviations of the correlations between GMD and age for all training dataset sizes split out over the three settings.

Firstly, as shown in Figure 5, when the centres had an equal mean age, the correlation between GMD and age was approximately equal when more than 20 subjects per centre were used to train the harmonization model. When 20 subjects or less per centre were used to train the model, we found a slight decrease in correlation as training dataset size decreased. Secondly, in the setting where two groups had a small age difference, the correlation between GMD and age was approximately the same when more than 20 subjects were used to train the harmonization model. However, we found a decrease in correlation between GMD and age when 20 subjects or less per centre were used to train the model. Additionally, when the training dataset contained 20 subjects or less per centre, we found a larger decrease in the correlation between GMD and age compared to when the two groups had an equal mean age, which is in accordance with our hypothesis. Finally, in the setting with a large age difference between the two groups, the correlation between GMD and age was approximately the same when more than 20 subjects were used to train the harmonization model. However, we again found a decrease in correlation between GMD and age when 20 subjects or less per were used to train the model. In addition, when the training dataset contained 20 subjects or less per centre, we found a larger decrease in correlation compared to both the setting where there was no age difference, and where there was a small age difference between the two groups, which is in accordance with our hypothesis. It should also be mentioned that, similar to the results of research question 1, we saw an increase in variance of the correlation between GMD and age when the training dataset contained fewer subjects. This increase in variance was stronger for settings with larger age differences between centres. Thus, in correspondence to the results of research question 1, the quality of MRI site harmonization becomes precarious when there is not enough data to train the ComBat model, and this effect is stronger when there are large age differences between sites. Descriptive statistics of the interaction effect of training dataset size and age differences between sites on MRI site harmonization quality are presented in Table 4.

Subsequently, a repeated measures ANOVA was performed to evaluate the interaction effect between training dataset size and size of mean age differences between centres on GMD. The repeated measures ANOVA revealed a significant but small interaction effect (F (14, 20979) = 19.71, $p < .001$, $\eta^2_g = .005$).

Table 4

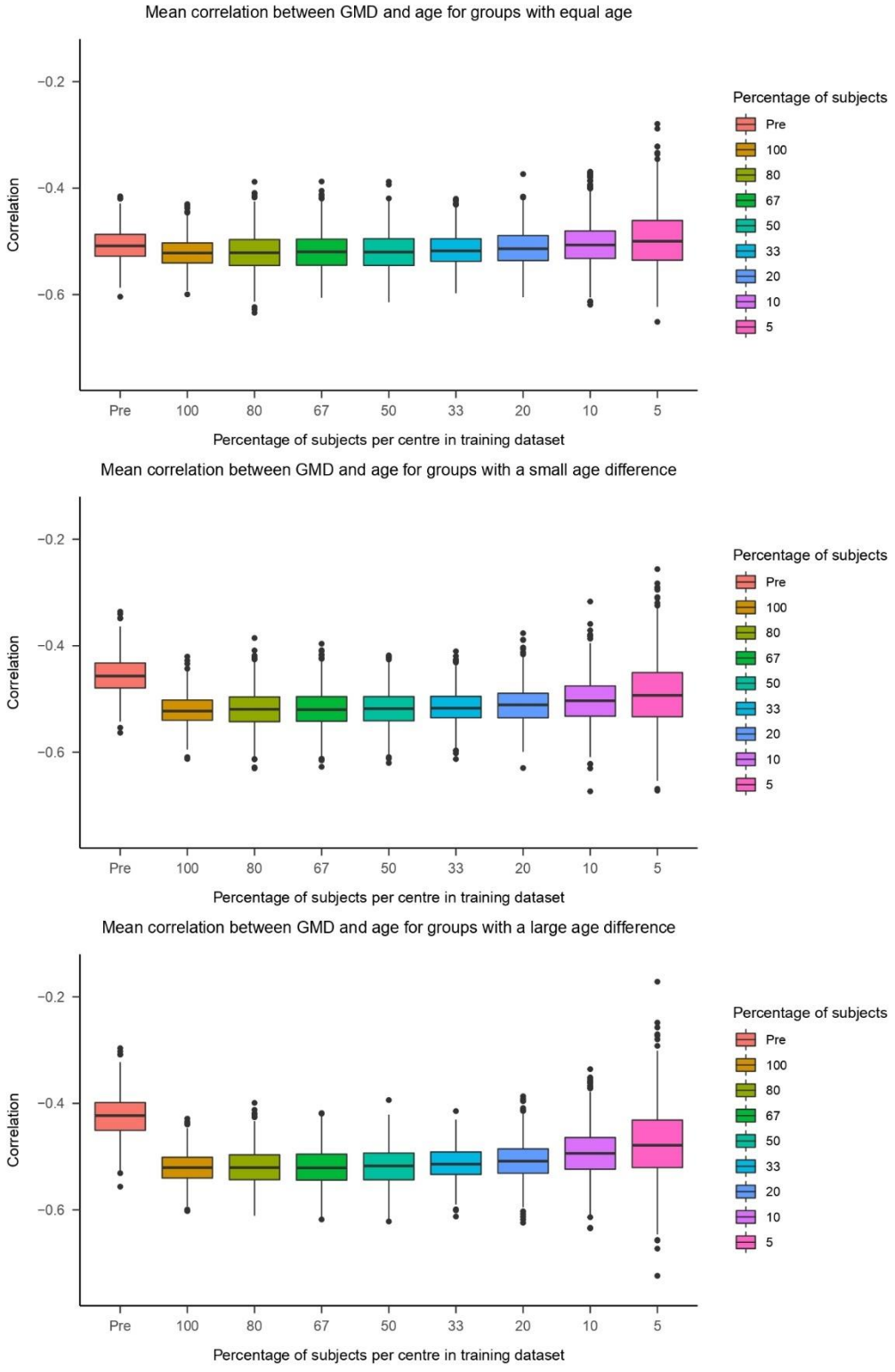
Descriptive statistics of the correlation between GMD and age for all training dataset sizes in a setting with no age difference, a small age difference and a large age difference.

	None		Small		Large	
	M	*SD*	*M*	*SD*	*M*	*SD*
Pre	-.507	.030	-.456	.033	-.423	.038
100	-.521	.028	-.521	.028	-.520	.028
80	-.520	.036	-.519	.034	-.520	.035
67	-.519	.036	-.519	.036	-.519	.035
50	-.519	.035	-.517	.035	-.518	.035
33	-.515	.031	-.516	.030	-.513	.031
20	-.512	.036	-.511	.033	-.507	.035
10	-.505	.042	-.502	.043	-.493	.046
5	-.496	.054	-.491	.063	-.476	.067

Note. "Pre" represents descriptive statistics for the unharmonized data.

Figure 5

The correlation between GMD and age over all repetitions for all training dataset sizes in a setting with groups having an equal age, a small age difference and a large age difference. The x-axis represents the number of subjects per centre in the training dataset. The y-axis represents the correlation between GMD and age.



Research question 3

To assess whether the effect of training dataset size on MRI site harmonization was dependent on the size of site differences, we created four settings where the data contained small, medium, large and huge site differences. These settings represent site effects with a partial η^2 of .01, .06, .14 and .26 respectively. Likewise, a site effect size of .26 was used in the analysis of research question 1. Figure 6 shows the correlation between GMD and age over all repetitions for all training dataset sizes, for the four settings.

Firstly, as can be deduced from Figure 6, when there was a small site effect, the correlation between GMD and age was approximately the same in the conditions where more than 10 subjects per centre were used to train the model. However, the correlation between GMD and age decreased when fewer than 10 subjects per centre were used. Another point that can be made is that the mean correlation between GMD and age for the unharmonized data is higher than all mean correlations between GMD and age for the harmonized data. This implies that when there are only small site effects, harmonization might not always be beneficial for data analysis.

Secondly, when there was a medium site effect, the correlation between GMD and age was approximately the same in the conditions where more than ten subjects per centre were used to train the model. In addition, we found an increase in variance when the training dataset contained only five subjects per centre. Similar to when there was a small site effect, the mean correlation between GMD and age for the unharmonized data was higher than the correlation between GMD and age when five subjects per centre were used to harmonize the data, implying that, when there is a medium site effect, harmonization could harm data analysis when the training dataset contains a small number of subjects per centre.

Thirdly, when there was a large site effect, the correlation between GMD and age was approximately the same in the conditions where more than ten subjects per centre were used to train the model. In addition, we found an increase in variance when the training dataset contained only five subjects per centre. As opposed to the settings with a small and medium site effect, in the setting of a large site effect, the mean correlation between GMD and age for the unharmonized data was lower than all mean correlations between GMD and age for the harmonized data. Thus, in a setting with a large site effect, data analyses benefit from the harmonization of data.

Finally, when there was a huge site effect, the correlation between GMD and age was approximately the same in the conditions where more than ten subjects per centre were used

to train the model. Again, we found an increase in variance when only five subjects per centre were used to train the model. Descriptive statistics of the interaction effect of training dataset size and site effect on MRI site harmonization quality are presented in Table 5

Following these analyses a repeated measures ANOVA was performed to compare the effect of training dataset size and site effect on GMD. A repeated measures ANOVA revealed that there was an interaction between the effects of training dataset size and size of site differences (F (21, 27972) = 10.56, $p < .001$, $\eta^2_g = .003$).

Table 5

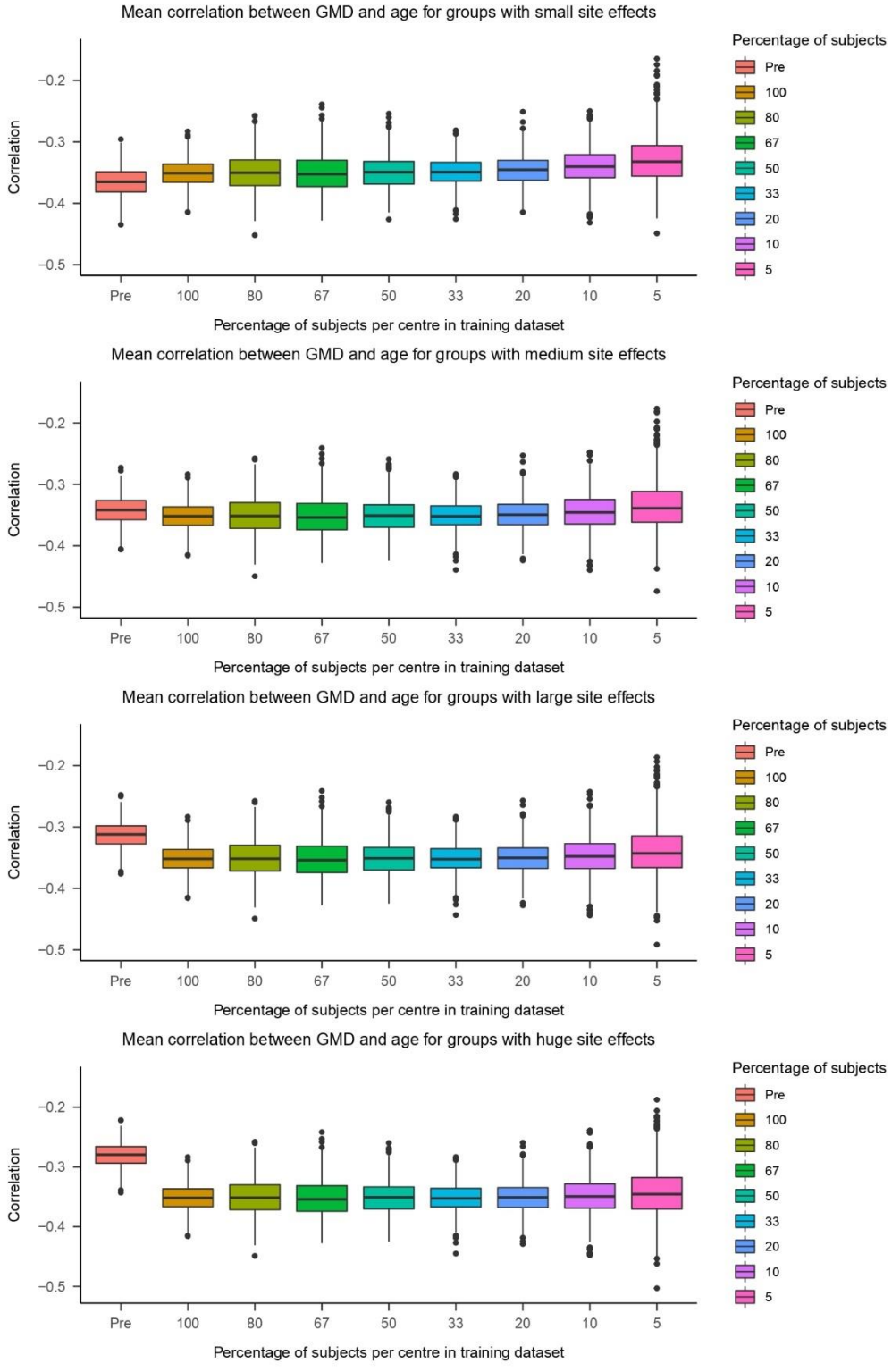
Descriptive statistics of the correlation between GMD and age for all training dataset sizes in a setting with a small, medium, large and huge site effect respectively.

	Small		Medium		Large		Huge	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pre	-.365	.023	-.341	.022	-.312	.021	-.280	.020
100	-.351	.021	-.351	.031	-.351	.031	-.351	.031
80	-.350	.036	-.351	.031	-.351	.031	-.351	.031
67	-.351	.031	-.352	.031	-.352	.031	-.352	.031
50	-.349	.027	-.350	.027	-.351	.027	-.351	.027
33	-.349	.024	-.351	.024	-.351	.024	-.352	.023
20	-.346	.024	-.349	.024	-.350	.025	-.351	.025
10	-.340	.029	-.344	.030	-.347	.030	-.349	.031
5	-.329	.040	-.335	.040	-.340	.042	-.343	.042

Note. “Pre” represents descriptive statistics for the unharmonized data.

Figure 6

The correlation between GMD and age over all repetitions for all training dataset sizes in a setting with are small, medium and large site effect respectively. The x-axis represents the number of subjects per centre in the training dataset. The y-axis represents the correlation between GMD and age.



Discussion

The aim of this thesis was to study the effect of training dataset size on MRI site harmonization. I hypothesized that the larger the size of the training dataset, the better the performance of MRI site harmonization. In addition, I studied whether the effect of training dataset size was dependent on age differences between sites, and the size of site effects. I hypothesized that the effect of training dataset size was larger when between-site differences in age were large. Also, I hypothesized that the effect of training dataset size was larger when site-related variance was large.

In summary, we found that harmonization had a positive effect on the correlation between GMD and age. However, we found no clear pattern in the relation between the quality of site harmonization and training dataset size in our empirical study. Therefore, our study confirms that MRI site harmonization, particularly ComBat site harmonization, has a positive effect on multi-site data, which is in line with previous research on this topic (Chen et al., 2011; Nan et al., 2022). Findings from our empirical study do not provide evidence that a larger number of subjects per centre in the training dataset translates to a higher quality of MRI site harmonization.

In our simulation study, we found no effect of training dataset size on the quality of site harmonization when the centre effect size was representative of neuroimaging research (i.e., $\eta^2 = .26$) and the mean age of subjects did not differ between centres. Apparently, the number of subjects per centre in the training dataset does not affect the quality of MRI site harmonization in that situation, which contradicts our hypothesis. We did, however, find an increase in variance of the quality of site harmonization when the number of subjects per centre in the training dataset was 20 or lower. This suggests that the quality of MRI site harmonization becomes precarious when there is not enough data to train the ComBat model, and this should therefore be taken into consideration when working with small datasets. In addition, we found that, for every repetition, and for every training dataset size, the quality of site harmonization increased after harmonization. This suggests that multi-site MRI data always benefits from harmonization, regardless of the size of the training dataset, given that the centre effect size is representative of neuroimaging research and the mean age of subjects does not differ between centres.

The conclusion as to whether the effect of training dataset size was dependent on age differences between sites is that the effect of training dataset size was larger when age differences were large, which is as hypothesized. Also, when the number of subjects per

centre in the training dataset was 20 or lower, we found a decrease in the quality of MRI site harmonization. These findings suggest that a training dataset of at least 20 subjects per centre is required to train the harmonization model when there are small differences between centres on covariates of which its effects overlap with the centre effect, and more subjects are required when these differences between centres are larger. An overlap between site variance and site differences on covariates increases model complexity, since it is more difficult for the harmonization model to separate the two sources of variance, and thus more data is required to train the model. The conclusion as to whether the effect of training dataset size was dependent on the size of site effects is that the effect of training dataset size was larger when site effects were small, which contradicts our hypothesis. However, we also found that when there are only small site effects, harmonization might not always be beneficial for data analysis, since harmonization in these cases impaired recovery of the anticipated effect.

This study makes several noteworthy contributions to the field of MRI site harmonization. Firstly, this study showed MRI site harmonization can be applied to small datasets. This is highly significant, as large multi-centre collaboratives such as the ENIGMA consortium and the ADNI mostly contain data from research centres with more than 20 subjects, which makes harmonization feasible according to our results. However, researchers have to take into account that smaller datasets come with more variable harmonization quality. This implies that in these situations, for some studies harmonization might be more successful than for other studies. Secondly, our findings showed that large differences on covariates, of which its effect overlaps with the centre effect, negatively affect the quality of MRI site harmonization for small datasets. Large differences do not only affect the quality of site harmonization, but also the variance of the quality of site harmonization. Thirdly, our results showed that large site effects have a negative influence on the quality of MRI site harmonization for smaller datasets. Again, larger site effects do not only influence the quality of site harmonization, but also its variance. Finally, our findings showed that when site effects are (too) small, MRI site harmonization, particularly harmonization using ComBat, has a negative effect on multi-centre datasets. Therefore, knowledge of the size of the batch effects in the data is crucial when deciding on whether to apply MRI site harmonization to the data.

The findings of this study have to be seen in light of some limitations. First, our empirical dataset apparently lacked the sufficient amount of power to detect small effects, which therefore makes the results from our empirical study less reliable. The results of the empirical study could also suggest that this particular harmonization model was simple enough to be accurately fitted using smaller parts of the data. Underfitting could also explain

why harmonization was unsuccessful in the small site effect condition. Second, the site effects that were present in the empirical dataset, and were thus present in the simulated datasets, have distorted the relation between GMD and age. Therefore, this study did not have a ground truth to compare our results to. Although the ground truth was missing, results from the empirical study give a general idea on the relation between GMD and age. Third, our data generation procedure was extremely laborious. Although a cross validation approach was necessary in our empirical study, our simulation study did not require this diligent approach. Instead of using a cross validation approach, both the training and test dataset could have been simulated from the same distribution, which would have saved us a considerable amount of time and effort. Since the datasets are independent in both approaches, results would likely not have been different. Finally, even though simulation studies come with significant benefits, such as the flexibility of altering variables to examine their effects and the reduction in costs and time efficiency, it is still difficult to simulate realistic data. Empirical data is subject to variance from multiple sources, which therefore makes the data extremely complex to simulate.

A challenging task for further research is to increase harmonization model complexity. This could be done by adding additional covariates to the model, which would reduce the model bias. Also, the variance of the covariates could be increased to increase model variance, and thus model complexity. However, altering model complexity would increase the risk of either underfitting or overfitting the harmonization model, which would therefore require larger datasets. Moreover, another promising line of research would be to carry out the same analyses on other MRI scanning techniques, such as Diffusion Tensor Imaging (DTI) and functional MRI (fMRI). Results from this study suggest that more subjects are required to successfully harmonize DTI images, as these images are generally subject to larger between scanner variation. Additionally, since fMRI images are generally subject to smaller between scanner variation, our results suggest that harmonization might not always be beneficial for multicentre fMRI studies.

In conclusion, this study showed that the size of the training dataset does not affect the quality of MRI site harmonization. However, an effect of training dataset size is present when the effect is dependent on age differences between sites, and on the size of site effects, when datasets are small. These results may aid future neuroimaging studies using scans from multiple scanning sites.

References

Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. In D. Banks, F. R. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, Clustering, and Data Mining Applications* (pp. 639–647). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17103-1_60

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384. https://doi.org/10.3758/bf03192707

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE*, *6*(2), e17238. https://doi.org/10.1371/journal.pone.0017238

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

DeBruine L. (2021). *faux: Simulation for Factorial Designs R package version 1.1.0. Zenodo.* http://doi.org/10.5281/zenodo.2669586

De Vos, F., Schouten, T. M., Koini, M., Bouts, M. J., Feis, R. A., Lechner, A., Schmidt, R., van Buchem, M. A., Verhey, F. R., Olde Rikkert, M. G., Scheltens, P., de Rooij, M., van der Grond, J., & Rombouts, S. A. (2020). Pre-trained MRI-based Alzheimer's disease classification models to classify memory clinic patients. *NeuroImage: Clinical*, *27*, 102303. https://doi.org/10.1016/j.nicl.2020.102303

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M. & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Gennatas, E. D., Avants, B. B., Wolf, D. H., Satterthwaite, T. D., Ruparel, K., Ciric, R., Hakonarson, H., Gur, R. E. & Gur, R. C. (2017). Age-Related Effects and Sex Differences in Gray Matter Density, Volume, Mass, and Cortical Thickness from Childhood to Young Adulthood. *The Journal of Neuroscience*, *37*(20), 5065–5073. https://doi.org/10.1523/jneurosci.3550-16.2017

Handels, R. L., Aalten, P., Wolfs, C. A., OldeRikkert, M., Scheltens, P., Visser, P. J., Joore, M. A., Severens, J. L., & Verhey, F. R. (2012). Diagnostic and economic evaluation of new biomarkers for Alzheimer's disease: the research protocol of a prospective cohort study. *BMC Neurology*, *12*(1). https://doi.org/10.1186/1471-2377-12-72

Jansen, W. J., Handels, R. L., Visser, P. J., Aalten, P., Bouwman, F., Claassen, J., van Domburg, P., Hoff, E., Hoogmoed, J., Leentjens, A. F., Rikkert, M. O., Oleksik, A. M., Smid, M., Scheltens, P., Wolfs, C., Verhey, F., & Ramakers, I. H. (2016). The Diagnostic and Prognostic Value of Neuropsychological Assessment in Memory Clinic Patients. *Journal of Alzheimer's Disease*, *55*(2), 679–689. https://doi.org/10.3233/jad-160126

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kassambara, A. (2020*). Rstatix: Pipe-friendly framework for basic statistical tests.* https://CRAN.R-project.org/package=rstatix

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Nan, Y., Ser, J. D., Walsh, S., Schönlieb, C., Roberts, M., Selby, I., Howard, K., Owen, J., Neville, J., Guiot, J., Ernst, B., Pastor, A., Alberich-Bayarri, A., Menzel, M. I., Walsh, S., Vos, W., Flerin, N., Charbonnier, J. P., Van Rikxoort, E., . . . Yang, G. (2022). Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. *Information Fusion*, *82*, 99–122. https://doi.org/10.1016/j.inffus.2022.01.001

Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P. J., Jeurissen, B., Ribbens, A., Den Dekker, A. J., & Sijbers, J. (2020). Harmonization of Brain Diffusion MRI: Concepts and Methods. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00396

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., Weickert, C. S., Weickert, T., Bruggemann, J., Kircher, T., Nenadić, I., Cairns, M. J., Seal, M., Schall, U., Henskens, F., Fullerton, J. M., Mowry, B., Pantelis, C., Lenroot, R., . . . Pineda-Zapata, J. (2020). Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage*, *218*, 116956. https://doi.org/10.1016/j.neuroimage.2020.116956

Ramanoël, S., Hoyau, E., Kauffmann, L., Renard, F., Pichat, C., Boudiaf, N., Krainik, A., Jaillard, A. & Baciu, M. (2018, 3 augustus). Gray Matter Volume and Cognitive Performance During Normal Aging. A Voxel-Based Morphometry Study. *Frontiers in Aging Neuroscience*, *10*. https://doi.org/10.3389/fnagi.2018.00235

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K.,

Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M. & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Zhu, A. H., Moyer, D. C., Nir, T. M., Thompson, P. M., & Jahanshad, N. (2019). Challenges and Opportunities in dMRI Data Harmonization. *Computational Diffusion MRI*, 157–172. https://doi.org/10.1007/978-3-030-05831-9_13