



OPEN ACCESS

EDITED BY

Hailong Song,
University of Pennsylvania, United States

REVIEWED BY

Gina Caetano,
University of Lisbon, Portugal
Daniel Brennan,
University of Pennsylvania, United States

*CORRESPONDENCE

N. G. Harris
✉ ngharris@ucla.edu

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 16 April 2025

ACCEPTED 11 July 2025

PUBLISHED 20 August 2025

CITATION

Kislik G, Fox R, Korotcov AV, Zhou J, Febo M, Moghadas B, Bibic A, Zou Y, Wan J, Koehler RC, Adebayo T, Burns MP, McCabe JT, Wang KK, Huie JR, Ferguson AR, Paydar A, Wanner IB, Harris NG and the TOP-NT Investigators (2025) Improved injury detection through harmonizing multi-site neuroimaging data after experimental TBI: a Translational Outcomes Project in Neurotrauma consortium study. *Front. Neurol.* 16:1612598. doi: 10.3389/fneur.2025.1612598

COPYRIGHT

© 2025 Kislik, Fox, Korotcov, Zhou, Febo, Moghadas, Bibic, Zou, Wan, Koehler, Adebayo, Burns, McCabe, Wang, Huie, Ferguson, Paydar, Wanner, Harris and the TOP-NT Investigators. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Improved injury detection through harmonizing multi-site neuroimaging data after experimental TBI: a Translational Outcomes Project in Neurotrauma consortium study

G. Kislik^{1†}, R. Fox^{1†}, A. V. Korotcov^{2,3}, J. Zhou⁴, M. Febo⁵, Babak Moghadas⁴, Adnan Bibic^{4,6}, Yunfan Zou⁴, Jieru Wan⁷, R. C. Koehler⁷, T. Adebayo^{2,3}, M. P. Burns⁸, J. T. McCabe⁹, K. K. Wang¹⁰, J. R. Huie¹¹, A. R. Ferguson¹¹, A. Paydar¹, I. B. Wanner¹², N. G. Harris^{1,12*} and the TOP-NT Investigators

¹UCLA Brain Injury Research Center, Department of Neurosurgery, Geffen Medical School, University of California at Los Angeles, Los Angeles, CA, United States, ²Department of Radiology & Bioengineering, Uniformed Services University of the Health Sciences, Bethesda, MD, United States, ³Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, United States, ⁴Department of Radiology, Johns Hopkins University, Baltimore, MD, United States, ⁵Department of Psychiatry, University of Florida, Gainesville, FL, United States, ⁶Hugo W. Moser Research Institute at Kennedy Krieger, Baltimore, MD, United States, ⁷Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University, Baltimore, MD, United States, ⁸Department of Neuroscience, Georgetown University Medical Center, Washington, DC, United States, ⁹Department of Anatomy, Physiology & Genetics, Uniformed Services University of the Health Sciences, Bethesda, MD, United States, ¹⁰Morehouse School of Medicine, Atlanta, GA, United States, ¹¹University of California San Francisco, San Francisco, CA, United States, ¹²Semel Institute for Neuroscience and Human Behavior, Intellectual Development and Disabilities Research Center, University of California at Los Angeles, Los Angeles, CA, United States

Multi-site neuroimaging studies have become increasingly common in order to generate larger samples of reproducible data to answer questions associated with smaller effect sizes. The data harmonization model NeuroCombat has been shown to remove site effects introduced by differences in site-related technical variance while maintaining group differences, yet its effect on improving statistical power in pre-clinical models of CNS disease is unclear. The present study examined fractional anisotropy data computed from diffusion weighted imaging data at 3 and 30 days post-controlled cortical impact injury from 184 adult rats across four sites as part of the Translational-Outcome-Project-in-Neurotrauma (TOP-NT) Consortium. Findings supported prior clinical reports that NeuroCombat fails to remove site effects in data containing a high proportion-of-outliers (>5%) and skewness, which introduced significant variation in non-outlier sites. After removal of one outlier site and harmonization using a pooled sham population, the data displayed an increase in effect size and group level effects ($p < 0.01$) in both univariate and voxel-level volumes of pathology. This was characterized by movement toward similar distributions in voxel measurements (Kolmogorov–Smirnov $p < <0.001$ to >0.01) and statistical power increases within the ipsilateral cortex. Harmonization improved statistical power and frequency of significant differences in areas with existing group differences, thus improving the ability to detect regions affected by injury rather than by other confounds. These findings indicate the utility of NeuroCombat in reproducible data collection, where biological differences can be accurately revealed to allow for greater reliability in multi-site neuroimaging studies.

KEYWORDS

diffusion-weighted imaging, traumatic brain injury, harmonization, multi-site, controlled cortical impact injury, axonal injury

Introduction

Diffusion-weighted imaging (DWI) is frequently used to study restricted water movement in brain tissue, and has been shown to be especially useful for investigating white matter pathology after traumatic brain injury (TBI) (1–3). Fitting DWI data to the tensor produces several metrics which describe the motion and direction of fluid, including the ratio of the primary eigenvalues—fractional anisotropy (FA) (4). Clinical multi-site studies using DWI are increasingly common, largely since the increased sample size results in improved statistical power (5, 6). New multi-center, preclinical studies are also beginning to be published (7, 8) so that it will be important to determine whether the application of tools and algorithms that are currently being used for clinical data harmonization can also be used preclinically. An important component of such multi-center studies is that they afford greater sensitivity to detect group differences, allowing hypotheses to be tested that have smaller effect sizes. These effects would otherwise not be detected because no single laboratory could easily produce the large sample size required. However, an unwanted consequence of combining data from multiple sites is the technical or non-biological variation that is introduced as a result of differences in equipment, especially scanner vendor and protocol type, among other factors (9). This undesired variation is referred to as site effects and has been a key target to improve the reproducibility of multi-site neuroimaging studies.

The development of multi-site studies has also come with an enhanced need to ensure data access and reproducibility. Noted in the FAIR principles which promote findability, accessibility, interoperability, and reusability, data integration both within and outside of consortia is key to ensuring good stewardship and accessibility to information for all stakeholders of a project (10). The growing size of data collected presents a further challenge for multi-site studies. Fluctuations in procedure, data analysis and collection, as well as instrumentation that result in site effects, can hinder replicability and the ability to compare data. When prospective alignment of procedures before data acquisition fails to enhance data harmonization, it becomes necessary to use retroactive methods to remove non-biological variation from data.

Numerous statistical methods have been created to harmonize neuroimaging data from multiple sites in order to remove site effects but preserve group differences (11). NeuroComBat was developed (12) as a modification of ComBat that allowed it to remove batch effects from neuroimaging, rather than genomic data sets (13). NeuroCombat (ComBat) has been shown to be able to consistently remove inter-site variation using an empirical Bayesian approach (12, 13). It has been employed to successfully remove site effects across several research locations and clinical protocols without obscuring biological variation (14, 15). It is notable, however, that in the absence of site effects, ComBat paradoxically *reduced* detectable biological variation among clinical data (16), indicating that some caution is required when applying this methodology.

While a large number of studies have demonstrated the effectiveness of ComBat in clinical populations, few have applied it to neuroimaging data from preclinical models of TBI. The Epilepsy Bioinformatics Study for Anti-epileptogenic Therapy (EpiBios4Rx) preclinical team previously demonstrated removal of magnetic field strength effects on FA values in the corpus callosum following ComBat harmonization, and this revealed significantly reduced FA in TBI rats (17). In that work, consistent areas of injury were demonstrated from data acquired across different scanners at the same site. Herein we were able to corroborate this, both across the four different research sites that were included in the Translational Outcome Project in Neurotrauma (TOP-NT), as well as across scanners with different field strengths. Unlike in previous work where the injuries were consistent across scanners (17), the severity of injury used in the current study was purposely varied to test for the utility of FA and other tensor-derived dependent variables as potential biomarkers predictive of outcome. We therefore took injury severity into account when creating the ComBat model, using the degree of whole brain atrophy as a proxy. Additionally, previous studies have demonstrated increased statistical power as a result of harmonization but have not examined voxel-level change in order to determine the potential effects on a viable injury site versus more remote regions (5, 6). Thus, in addition to assessing whole-brain level univariate volumes of pathology across groups, we also assessed ComBat-related changes in statistical power, effect size and variability at the whole-brain level across individual *voxels* between groups (Figure 1). These measures, in conjunction with univariate harmonization, allowed us to investigate the ideal level at which ComBat harmonization could be applied to preclinical brain injury neuroimaging data, as well as to evaluate brain regional changes as a result of harmonization. Novel here is that we also applied ComBat-seq to harmonize univariate metrics derived from neuroimaging data. Although ComBat-seq was originally designed to harmonize RNA sequencing data (18), univariate data derived from the number of voxels is a suitable input. This study is important for delineating how NeuroCombat affects statistical power to detect injury across the whole brain, as well as to quantify changes in effect size after harmonization. This will improve the scientific community's understanding of NeuroCombat's capabilities when applied to pre-clinical datasets. Additionally, it will allow for the characterization of brain areas which benefit from harmonization and to better develop appropriate use-cases for harmonization.

Methods

Experimental protocol

As part of the TOP-NT consortium project, there were four data acquisition sites: the University of California Los Angeles (UCLA), the University of Florida and Morehouse School of Medicine (UF/MSM), Georgetown University and Uniformed Services University (GU/USU), and John Hopkins University (JHU), and a data analytics

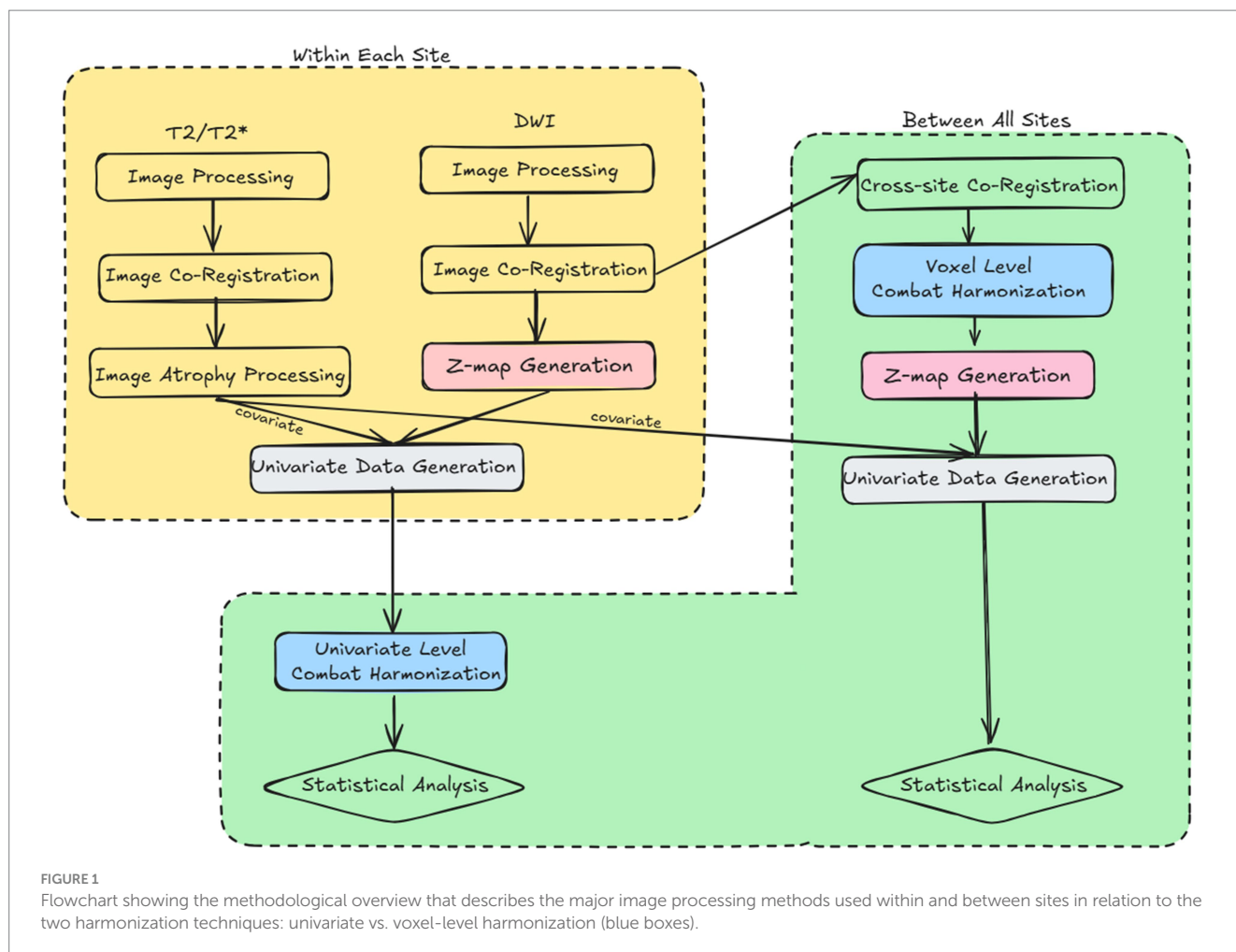


FIGURE 1

Flowchart showing the methodological overview that describes the major image processing methods used within and between sites in relation to the two harmonization techniques: univariate vs. voxel-level harmonization (blue boxes).

site—University of California San Francisco (UCSF). All study protocols used were in compliance with the Public Health Service Policy on Humane Care and Use of Laboratory Animals and approved by each site-specific Animal Research Committee, UCLA Approval #2021–082; UF/MSM Approval # 202010145; GU/USU Approval APG-21-044; JHU Approval #RA22M370. A total of 186 adult male and female, Sprague Dawley rats were used for this study and were randomly assigned to each group (UCLA: 33 TBI, 17 sham, JHU: 35 TBI, 14 sham, GU/USU: 33 TBI, 16 sham, UF/MSM: 27 TBI, 11 sham). Diffusion imaging data were acquired at 3 and 30 days after either controlled cortical impact (CCI) injury or sham anesthetic controls in adult, male (Mean \pm SD body weight 317 ± 43.7 g, 271.3 ± 20.3 g, 225 ± 14.8 g, 369 ± 16.1 g, at UCLA, GU/USU, JHU and UF/MSM, respectively) and female rats (body weight 254 ± 19.1 g, 239.3 ± 13.8 g, 183 ± 12.7 g, 256 ± 7.9 g, at UCLA, GU/USU, JHU and UF/MSM, respectively). A total of 343 scans were collected across the two post-injury times across all sites (25 datasets were missing). Prior to data acquisition, the injury and imaging protocols were harmonized across the sites using standard operating procedures (7, 19). Post-acquisition data analysis, including all pre- and post-processing data steps were also harmonized using a BASH script that wrapped multiple commands which was shared across the sites. The detailed methods below document the cross-site harmonized methods and any differences across sites are noted. All data were acquired with adherence to the ARRIVE guidelines (20) including group assignment randomization, blinding by use of unique animal IDs across site with

semi-automated analysis routines, as well as conforming to the other eight guidelines on protocol, analysis and reporting standards.

Controlled cortical impact

Rats were housed in pairs before and after the surgery in standard cages with ad libitum access to water and food. Rats were maintained on a 12:12 h light/dark cycle in the vivarium. Estrus cycle was assessed for female rats by vaginal lavage/swab at the time of surgery to determine the stage of the estrus cycle of the rat on the injury day. Two different levels of CCI were induced at each site by varying the deformation depth below the dura of the impactor using similar methods across the sites (Table 1). Rats were anesthetized with 5% isoflurane vaporized in 1 L/min O₂ (2% isoflurane was used for maintenance of the anesthesia with adjustments between 1.5 and 2%), and then were head-fixed in a stereotactic frame and their body temperature was maintained at 37°C using a temperature-controlled heating-pad. After prophylactic administration of analgesics according to the site-specific animal protocol, aseptic technique was used to make a midline, longitudinal skin incision, followed by a 6 mm craniectomy (5 mm UF/MSM) centered at +3.5 mm anteroposterior and 3 mm lateral to the midline over the left hemisphere using a dental drill cooled under intermittent saline. CCI injury was conducted using a 5 mm (4 mm UF/MSM) diameter, bevel-edge, metal impactor tip at an

TABLE 1 Injury parameters across the sites.

Parameter	UCLA	GT/USU	JHU	UF/MSM
Craniectomy size (mm)	6	6 to 7	6	4
Antero-posterior position (mm)	−3.5	−3.5	−3.5	−2.5
Medial-lateral position (mm)	3.5	3.0	3.0	3.0
Hemisphere	left	left	left	right
Tip diameter (mm)	5	5	5	4
Injury device	Leica one	Leica one	Leica one	Leica one
Tip material	Metal	Metal	Metal	Metal
Tip shape	Flat	Round	Flat	Flat
Angle of impactor (degrees)	13–18	12	5–10	0
Dwell time (ms)	200	200	200	200
Speed (m/s)	5	5	5	5
Depth below dura (mm)	1.7; 2.1	1.7; 2.1; 2.8	1.0; 1.2; 1.5; 1.7	1.7; 2.1
Angle to perpendicular (degrees)	10–15	5–10	5–10	0
Avg. time under anaesthesia (means ± STD; min)	39 ± 0.01	45.8 ± 12.7	25.1 ± 1.5	-

angle parallel to the dural surface and at impact depth below dura, speeds and dwell times ranging from: 1.3–2.3 mm, 3.5–6 m/s and 100–240 ms, using an electromagnetic impact system (Leica Biosystems, IL, United States). The craniectomy was closed with a non-toxic sealant (Kwik-Cast, WP instruments, United States) and the skin sutured.

Image acquisition

Data were acquired on Bruker Biospec consoles (Bruker, Billerica, MA, United States) connected to magnets from a variety of vendors and with different field strengths (Tesla, T) across the sites (7 T -UCLA and GU/USU, 11.7 T – JHU and 11.1 T—UF/MSM) with gradients of a varying peak strength and rise time resulting in different slew rates. Radiofrequency coils were similar across three sites (birdcage volume transmit coil decoupled from a receive-only 4-channel phased-array surface coil) and a transceive quadrature coil at UF/MSM. Rats were imaged prone under isoflurane sedation (1–1.5%) vaporized in medical air (0.5 L/min) positioned in a cradle with head stabilization using a 3-point head fixation. The diffusion data were acquired after a number of other scans within the same imaging session as part of a larger protocol, taking around 60 min prior to diffusion imaging. The prior scans for functional and structural imaging were conducted under a mixture of continuous dexmedetomidine sedation (0.05 mg/kg/h, subcutaneously) and isoflurane (0.5%) prior to switching to isoflurane alone. Temperature was maintained at 37°C via homeothermic-controlled external air or water heating.

A 3-dimensional, single-shot, spin echo, echo planar imaging sequence was used to acquire diffusion-weighted images with directionally encoded, monopolar diffusion gradients applied along 88 different, non-collinear directions, with b values of 0, 1,000 and 3,000 s/mm² ($n = 4, 44$ and 44 , respectively), and no averaging. Four left–right, phase-reversed, $b = 0$ images were also acquired. Repetition and echo times varied slightly according to the supported gradient hardware (Table 2). Similarly, the diffusion gradient amplitude (big delta) was varied across site to arrive at similar b values while maintaining similar diffusion times, the time between application of the

diffusion-sensitizing gradients (Table 2). The acquisition data matrix was constrained to a 72x49x96 matrix size, in the 1st phase-encoding, 2nd phase-encoding and read-out directions, within a field of view (FOV) of 18x12.25x24mm, respectively in the dorsal-ventral, left–right and antero-posterior planes, resulting in an isotropic resolution of 250 μ m. Outer-volume signal suppression using multiple, 5–10 mm saturation slices were applied to reduce signal from outside the brain.

Whole brain anatomical data were acquired using a 3D-multi-gradient echo sequence with a variable matrix size and field-of-view across site, resulting in the same isotropic resolution of 160 μ m (Table 2). Eleven echoes were acquired from 2.8–51.8 ms with an echo spacing of 4.9 ms, a repetition time of 45–125 ms and a 13–20 flip angle (Table 2).

Image preprocessing

Bruker raw data were converted to NIFTI (21), brain extracted (22), denoised (23), and corrected for Gibbs ringing (24), after which phase distortions were unwrapped with TOPUP (25, 26) and Eddycorrect (27), all of which were implemented under MRTRIX (28). Data were then fit to the tensor to derive scalar images- fractional anisotropy (FA), axial, radial and mean diffusivity (AD, RD, MD, respectively). FA data were then used to derive an unbiased mean deformation template (MDT) at each site using successive registrations of rigid, affine and non-linear alignment under ANTS (29). The resulting affine transformations and warp fields were applied to all scalar data to align them to the local site MDT. Voxel-wise sham mean and standard deviation (SD) maps were computed for each tensor indices, which were then used together with each co-registered brain volume scalar to calculate regions of high and low indices relative to local sham data at a corrected z-score threshold based on a distribution-corrected z-score (30, 31) beginning with a desired z value of $|z| > \pm 3.1$ ($p < 0.001$). This DisCo-Z correction is required to adjust for group-wise bias originating for differences in sample size (30). The resulting corrected z values for thresholding the injury groups at each site at 03d/30d was: 3.85/3.75, 3.99/3.85, 4.19/4.19 and 5.08/5.08, respectively at UCLA, GU/USU, JHU, UF/MSM. For the sham group the adjusted

TABLE 2 MRI parameters for DWI and MGE parameter sequences across the sites.

Mode	UCLA (7T)	GU/USU (7T)	JHU (11.7T)	UF/MSM (11.1T)
	3D-SE-EPI, SingleShot	3D-SE-EPI, SingleShot	3D-SE-EPI, SingleShot	3D-SE-EPI, SingleShot
DWI				
FOV (mm)	24x18x12.25	24x18x12.25	24x18x12.25	24x18x12.25
Matrix Size (#voxels)	96x72x49	96x72x49	96x72x49	96x72x49
Resolution (um)	250	250	250	250
TR (ms)	1,000	1,000	1,000	1,000
TE (ms)	28.4	30	27.85	27.85
Flip angle (degrees)	90	90	90	90
BandWidth (KHz)	357	300	357	357
Big delta (ms)	12	12	12	12
LittleDelta (ms)	5.3	5	5.35	5.35
Bvalue1 (S/mm ²)	1,000	1,000	1,000	1,000
Bvalue2 (S/mm ²)	2,800	3,000	3,000	3,000

Mode	3D	3D	3D	3D
MGE				
FOV (mm)	25.6x19.52x12.8	32x32x16	25.6x25.6x19.2	25.6x25.6x19.2
Matrix size (#voxels)	160x122x80	200x200x100	160x160x120	160x160x120
Resolution (um)	160	160	160	160
TR (ms)	125	45	45	45
TE range (ms)	2.8–51.8	2.5–26.5	2.5–26.5	2.5–26.5
Echo spacing (ms)	4.9	3	3	3
Number echoes	11	9	9	9
Flip angle (degrees)	20	13	13	13
BandWidth (KHz)	100	100	100	100

SE-EPI = spin echo, echo-planar imaging.

z values at 03d/30d was: 2.67/2.72, 2.61/2.67, 2.52/2.52 and 2.38/2.38, respectively at UCLA, GU/USU, JHU, UF/MSM. The brain-wide volumes of tissue surviving these upper and lower z-score thresholds for each tensor scalar, herein designated $Scalar_{LOW}$ and $Scalar_{HIGH}$ volumes, were then quantified for each brain by counting the voxels that were greater or less than the z-score. For voxel-level harmonization, a multi-site MDT was calculated using the site-specific MDTs as input, and then each site-specific subject space data were transformed to the mean site MDT in a single step through application of the original warp and affine transformation together with those generated from site MDT to multi-site MDT space. Univariate harmonization statistics (e.g., corrected z-scores) were recalculated in the multi-site MDT space using the same method that had been applied at the site-specific level. Tensor-Based Deformation Analysis: Anatomical data were used to assess local tissue deformations to approximate tissue atrophy/compression and swelling/expansion as we have before (32, 33). Briefly, a mean deformation brain template was constructed from all data using ANTS (34) and the Jacobian Determinant was calculated from the resulting warp transformation fields. Sham Jacobian data were used to calculate voxel-based mean and standard deviation maps which were then used to calculate z statistic maps for all injured data. The volume of tissue surviving a threshold of $z \pm 3.01$ ($p < 0.01$, uncorrected) was determined for each injured brain and used to assess the region of local tissue contraction and referred to herein as tissue atrophy.

Harmonization and statistics

The R package NeuroCombat_1.0.13 was used for harmonization.¹ The covariates included in the model were day post-injury, atrophy, group, and sex (Figure 1). Site was used as the batch parameter. Given prior successes in applying NeuroCombat to cross-sectional data (scans less than 180 days apart) (16, 35) and its ability to adjust for covariates (36), all scans across days and groups were combined. By using large amounts of data in the model (samples >20), we hoped to replicate prior successes correcting for site effects in site-to-site traveling subject data (36). Atrophy was used as a proxy for the level of injury rather than the surgeon-designated mild/moderate scale. This is because designated injury severity is still not well modelled. Using atrophy as a continuous variable also afforded greater sensitivity to differences compared to a binary choice of gross injury severity. Days post injury were defined as the number of days since anesthesia in shams, and this was treated as a continuous variable. Empirical Bayes was left enabled as per the default NeuroCombat settings. All statistical analyses were conducted in RStudio version 2023.6.0.421 (Posit Team 2023). Cohen's d was calculated as a measure of effect size by dividing the difference between the mean injured and mean sham animal

¹ https://github.com/Jfortin1/neuroCombat_Rpackage

by the pooled standard deviation. The False Discovery Rate (FDR) correction was applied to correct for multiple comparisons of all statistics.

The number of brain voxels with FA greater or less than a statistical theory-derived distribution-corrected z-score (equivalent to $p < 0.001$ relative to site-specific sham data, Figure 2) are referred to as FA_{High} and FA_{Low} whole-brain volumes of pathology, respectively. We have operationally defined “whole-brain volume of pathology” based upon a voxel tensor scalar value being lower or greater than the z-score threshold, as before (31). It should be noted that variability among measurements in shams is not due to injury, but other likely sources of variance, such as between scanner effects and animal-to-animal differences. Relevant code to reproduce the analyses in this study can be found at the following link: https://github.com/ngharris/neurocombat_TBI.

Univariate Harmonization: The volume of brain regions significantly different from sham was derived from harmonization of univariate volume data derived from subject data co-registered to site-specific brain template space and is herein referred to as univariate-level harmonization (Figure 1). Ggplot2 was used for plotting violin plots for univariate data (37). Four-site data was

visualized with residual plots of values predicted by a regression model generated from the harmonized data across each site to demonstrate the presence of unequal variance among sites post-harmonization. As described above, NeuroCombat harmonization used atrophy, day post-injury, group and sex as covariates. Site was used as the batch parameter.

Since the data obtained from brain volumes used for univariate harmonization are count data derived from the number of voxels above or below a z-threshold, we also implemented Combat-seq. Combat-seq was originally designed for RNA sequencing count data, but can be natively applied to neuroimaging count data (18). We used sex, atrophy, and day post injury as covariates and treatment as the group variable. Site was used as the batch parameter. Bayes shrinkage remained disabled. Due to non-normality in the data, univariate metrics were compared using Kruskal Wallis tests to test for effect of group and site, and were FDR corrected for multiple comparisons. The estimation stats package was used to visualize effect size (38). p -values less than 0.05 were considered statistically significant for univariate statistics. Group differences were also maintained after harmonizing AD, MD, and RD data.

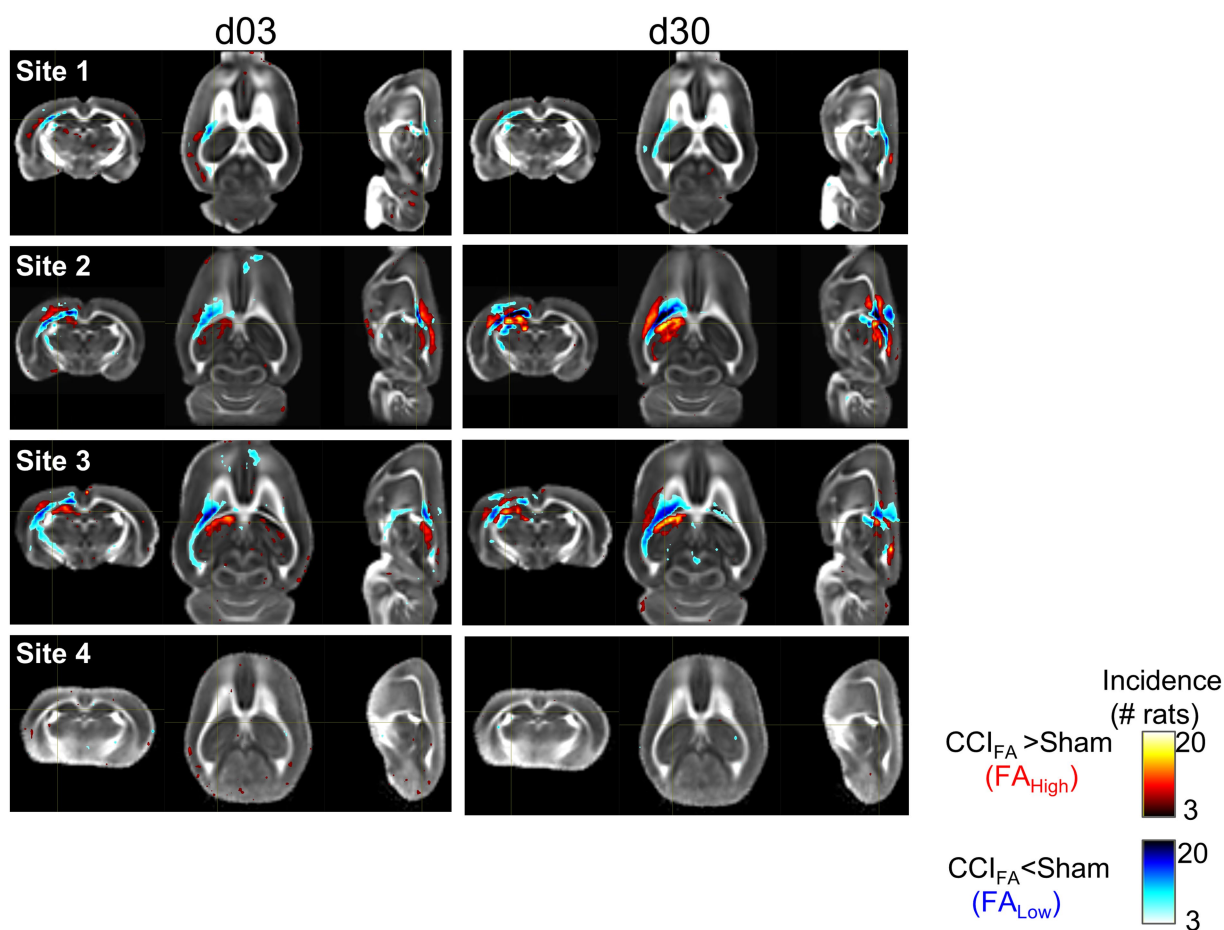


FIGURE 2

Voxel overlap maps of regions of FA-related pathology prior to harmonization at each site at three and 30d post-injury. Individual FA volume data from each injured rat that survived below the FA_{Low} or above the FA_{High} corrected zmap-threshold (red, blue colors respectively, $p < 0.001$) relative to local sham data are plotted as binary incidence maps for each site at day 3 and 30 post-injury. The voxel pseudocolors represent the overlap between the number of rats where FA survived threshold at each brain voxel location. Site 4 shows low overlap between rats.

Voxel level harmonization

Whole-brain volume of pathology data were also obtained after harmonization of whole-brain scalar data at the brain voxel level. This was derived from data co-registered to a multi-site-derived template, and is herein referred to as voxel-level harmonization (Figure 1). Loading of NIFTI files and generation of whole-brain maps of changes in effect size, standard deviation and power were conducted with the `oro.nifti` and `neurobase` packages (39, 40). As described above, NeuroCombat harmonization used atrophy, day post-injury, group and sex as covariates. Site was used as the batch parameter. Data points for voxels exceeding the range of 1.5 times the interquartile range (IQR) from the first (Q1) or third (Q3) quartiles of the whole dataset were identified as outliers. `FSLeys` was used to visualize effect size and power maps (41). Bland Altman plots were produced by plotting the difference of each paired voxel measurement against the two voxels' average value for each combination of site. Kolmogorov–Smirnov analysis tested for site effects and demonstrated differences in distributions of voxel-wise data before and after harmonization. Voxel overlap maps were used to demonstrate the proportions of rats with measurements significantly different from sham after harmonization, which displayed areas of biological injury. Frequency maps were generated by finding the number of rats different from mean sham at each voxel using a modified z-test (the quantity of the test value minus the sham mean at that voxel divided by the standard deviation of shams). Power at each voxel was calculated with the `R` `pwr` library's `pwr.t2n.test` function (42). No spatial smoothing or regularization was performed after harmonization. p values less than 0.01 were considered significant for voxel-level statistics.

Results

For clarity, we limited the reporting of data harmonization to fractional anisotropy, but we have included data on the other tensor scalars MD, AD and RD within the [Supplementary Figures S1–S15](#). These data were in general, qualitatively similar to the FA results described herein. Volume of atrophy was included as a covariate in harmonization to account for intentional variation in injury severity, but these volumes were not significantly different between sites (one-way ANOVA, $p > 0.05$).

Two harmonization methods were used. In the first, the volume of whole-brain pathology was derived from harmonization of univariate volume data derived from subject data co-registered to site-specific brain template space, and is herein referred to as univariate-level harmonization. In the second method, the volume of pathology was derived after harmonization of whole-brain FA data at the brain voxel level, derived from data coregistered to a multi-site-derived template, and is herein referred to as voxel-level harmonization.

Outlier sites lead to increased variance in non-outlier sites after univariate-level harmonization

Previous work (43) has shown that if the proportion of outliers is higher than 5% of the data, or if extreme outliers exist, this can severely distort the ability of NeuroCombat to remove site effects. In the current data, we found that the variance for both high and low FA

values was significantly increased within Site 4 sham data as compared to all other sites for univariate data before harmonization ($p < 0.001$). This increased variance within site 4 data was maintained even after harmonization when compared to the other three sites ($p < 0.001$). Consistent with this data, the residuals of FA values for sham rats predicted by a regression model generated with harmonized sham data showed unequal distributions with the largest spread in site 4 data. One potential cause of this heteroscedasticity could be attributed to the presence of outliers, which could lead to ineffective harmonization if all sites were included (Figure 3). However, log transformation of these data revealed similar results with a greater number of univariate outliers in site 4 compared to all and significantly increased variance. Removal of site 4 data resulted in an absence of site effects in the harmonized data ($p > 0.05$), further substantiating these findings. As a result, all further univariate analyses were conducted with data from only three sites in order to more accurately determine the utility of using NeuroCombat to harmonize preclinical imaging data from multiple sites. Additional analyses using data from all four sites can be found in [Supplementary Figures S16–S18](#).

Univariate-level harmonization with pooled but not site-specific shams leads to improved detection of injury

An important consideration when investigating differences in multi-site data is whether to pool shams from all sites, or whether to treat site-specific sham data independently for the calculation of univariate volume data. Pooled shams were used to generate the z-scored FA_{Low} and FA_{High} values for each rodent prior to harmonization. To gauge how effectively differences between sham and TBI rats were detected when considering pooled sham data versus sham data specific to each site, we examined FA_{Low} and FA_{High} distributions across the sites. Compared to unharmonized data, harmonization with pooled shams led to a 32% increase in the number of injured rats across all sites, where FA_{Low} volumes of pathology were more than two standard deviations larger than the mean sham animal across all sites ($n = 21$ vs. 30 animals at day 3 and $n = 19$ vs. 23 at day 30, respectively). The number of injured rats with FA_{High} volumes above the mean after harmonization were reduced or remained similar to unharmonized data ($n = 9$ vs. 8 at day 3, respectively and no change at day 30).

Harmonization with site-specific sham data afforded a lower improvement compared to the prior reported pooled sham harmonization; the number of injured animals with FA_{Low} volumes above sham mean data after harmonization increased by only 10% ($n = 33$ vs. 32 at day 3 and $n = 52$ vs. 45 at day 30, respectively). This aligns with previous suggestions that “mega-analyses,” in which pooling is done prior to statistical analyses, results in an improved ability to detect injury compared to meta-analyses that do not pool data (12). However, the number of injured rats with significantly different FA_{High} volumes after harmonization increased by 17.6% ($n = 11$ vs. 9 at day 3 and $n = 9$ vs. 8 at day 30, respectively). While univariate-level harmonization led to mixed results overall, the greatest improvement occurred in injured rat FA_{Low} volumes using pooled sham data. As a result, since NeuroCombat is intended for use on data pooled from multiple sites, pooled shams were used for all further univariate analyses. Pooled shams were also used for voxel-level analyses for consistency.

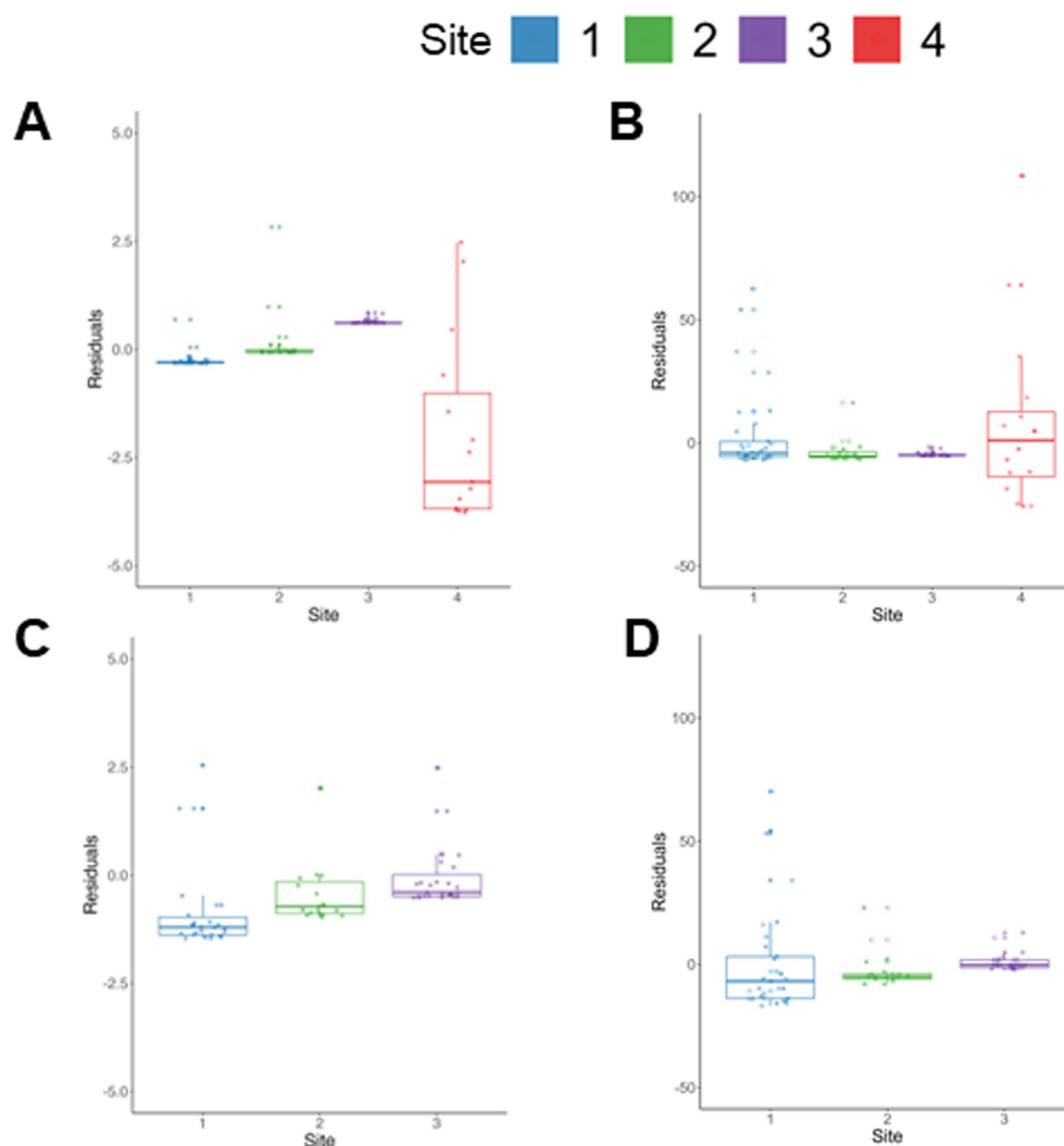


FIGURE 3

Impact of using four versus three site data for univariate-level harmonization on sham group data. The residuals of (A) FA_{Low} and (B) FA_{High} predicted using a model generated from univariate 4-site data harmonized with Combat-seq demonstrates a larger spread of residuals in site 4 and with fewer points centered around 0 compared to the other 3 sites. The residuals of (C) FA_{Low} and (D) FA_{High} predicted using a model generated from univariate 3-site data harmonized with Combat-seq indicate fewer outlier points across all 3 sites compared to the 4 site data.

Group differences after univariate data harmonization using pooled shams

Previous reports have demonstrated that removal of site-specific effects after harmonization would indicate that any remaining group differences could accurately be attributed to a true difference in pathology, thus demonstrating successful harmonization (12). Within the original, unharmonized, univariate data, there were significant sham vs. TBI group effects for FA_{Low} volume day 3 data (Figure 4A, $p < 0.001$, FDR correction, Kruskal-Wallis test) and significant site effects for FA_{High} volume at both time points ($p < 0.001$, Figures 4B,D). After univariate-level harmonization with Combat-seq, the group difference in FA_{Low} volume originally found at day 3 was retained, and a group difference at day 30 was introduced ($p < 0.001$, Figures 4A,C).

Finally, univariate-level harmonization failed to correct for the site effects observed in FA_{High} volumes at day 3 and 30 post-injury before harmonization (Figures 4B,D) where sham data from Site 1 had substantially higher volumes of pathology and were more variable than the other two sites ($p < 0.001$, Kruskal-Wallis test). This result was similarly found in the other measures, where group differences were maintained post-harmonization, but a significant effect of site was no longer found (Supplementary Figures S1, S6, S11).

Despite these unexpected effects, we calculated the effect size for each site in order to quantify any improvement in the ability to discriminate between experimental groups due to harmonization. Using the difference in univariate FA volume for injured versus pooled sham data at each timepoint, effect size changed in different directions for day 3 FA_{Low} data after harmonization at sites 1, 2 and 3 by -4.3 ,

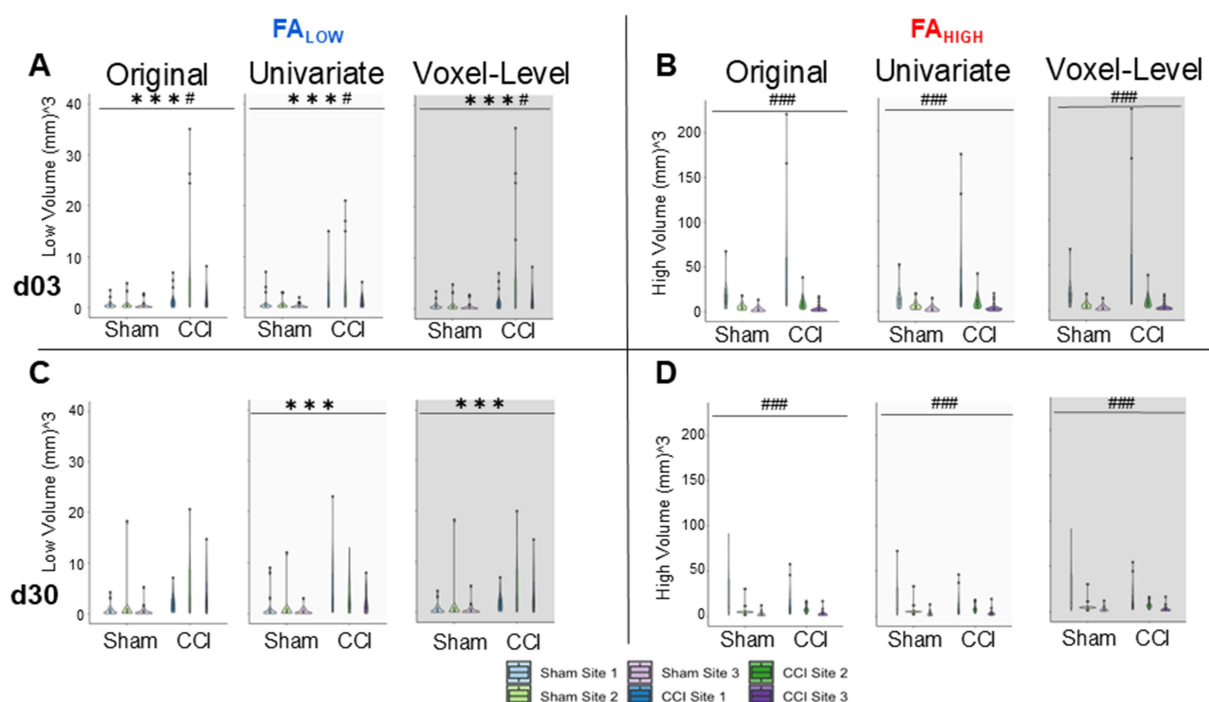


FIGURE 4
Effect of univariate and voxel-level harmonization on the burden of whole brain injury volume using pooled sham data. Brain volumes of pathology identified by (A,C) FA_{Low} and (B,D) FA_{High} at 3d (A,B) and 30d (C,D) are plotted for each site before ("original") and after univariate and voxel-level harmonization. ###/*** = a significant effect of site and group, respectively, $p < 0.001$, ##/** = $p < 0.01$, #/* = $p < 0.05$ (Kruskal-Wallis test). Harmonization for the univariate plots was performed with Combat-seq because the unharmonized data was count distributed.

1.35 and 2.1-fold, respectively. This was significantly different from zero ($p < 0.05$) for Sites 2 and 3 compared to pre-harmonization, when it was not previously different (Figure 5A). Similar but smaller effects were found for day 30 data at all sites (Figure 5B). Despite the introduction of group differences in the FA_{High} data at day 3 through harmonization, the magnitude of the effect size decreased for site 3 ($p < 0.05$ pre-harmonization, $p > 0.05$ after harmonization, Figure 5C) and it remained similar for other sites ($p > 0.05$).

Unexpectedly, neither Combat-seq or NeuroCombat were able to completely resolve the site effects in univariate FA_{Low} data on day 3, however, Combat-seq was able to uncover strong group effects at day 30 (Figures 4A,C; data not shown for NeuroCombat). NeuroCombat was also less sensitive for detecting FA_{Low} injury compared to Combat-seq, but similar for FA_{High} (data not shown for NeuroCombat). This occurred at both the day 3 and day 30 timepoints for FA_{Low} where Combat-Seq showed a more significant effect of group compared to NeuroCombat ($p < 0.001$ versus $p < 0.05$, respectively, Kruskal Wallis test). Group differences after harmonizing AD, MD, and RD data using Combat-seq were maintained with both NeuroCombat and Combat-seq.

Voxel-level harmonization leads to a common distribution of FA values and effect sizes across sites

By visual inspection of the group-level incidence maps of volumes of pathology described by FA_{Low} and FA_{High} , we found there were marked differences at site 4 compared to Sites 1–3 at both 3 and

30 days after injury (Figure 2), in agreement with the prior outlier volume analysis above. We confirmed this quantitatively for FA values; of the four sites analyzed, 10.5% of the FA values from Site 4 were found to lie more than 1.5 interquartile ranges outside the interquartile range of FA values when compared to the other three sites at each voxel location. The outliers among Sites 1, 2 and 3 were smaller, containing only 4.6, 1.7 and 0.9% outliers, respectively, which was more than 2-fold less than Site 4. Additionally, Site 4 FA population data were found to have an excess kurtosis of 0.497, compared to -0.774 , 0.159 and 0.274 in Sites 1, 2, and 3, respectively indicating a quite different distribution of FA volumes. These discrepancies were similar to the findings at the univariate volume level described in the prior sections. Harmonization of whole brain data resulted in large decreases in the standard deviation of Site 4 FA data and corresponding increases in variance in Sites 2 and 3, as well as to a lesser extent in Site 1 (Figure 6A). Since the presence of an outlier site severely biased the performance of NeuroCombat, further analysis of whole brain data was done with only Sites 1, 2 and 3.

Bland–Altman plots were used to quantify the effect of harmonization across the whole brain between sites at the voxel level across all subjects (Figures 6B,C). These graphs plot the mean value of the same voxel location from each pair of sites against their difference, and they were generated for each pair of sites before and after harmonization. In the original, unharmonized data, large differences were observed for all combinations of sites indicating site-related differences. After NeuroCombat harmonization, the difference between voxel measures between pairs of sites approached zero, indicating that FA values were more similar at the same brain location

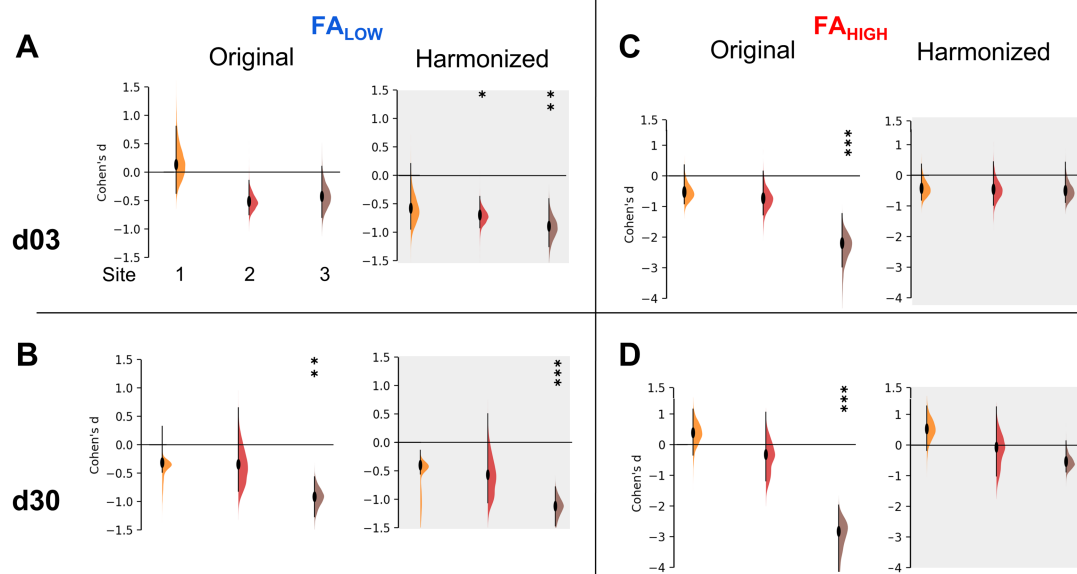


FIGURE 5

Effect sizes after univariate-level harmonization. Injury effect size measured by Cohen's D for each site before and after univariate harmonization for (A,B) FA_{Low} volumes and (C,D) FA_{High} volumes at (A,C) d03 and (B,D) d30 post-injury. The data demonstrates an increased effect size for all sites for FA_{Low} volumes, at 3 and 30d (A,B) but no change or a removal of significant effect size across sites for FA_{High} volumes (C,D). *** = an effect size significantly greater than 0, $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$ (t-test).

across sites after harmonization. This was supported by a Kolmogorov–Smirnov analysis, which showed significantly different site effects before harmonization ($p < 0.001$) and a resolution of these site effects after harmonization on average between each combination of sites ($p > 0.05$, Figures 6B,C).

Crucially, all group level effects shown prior to harmonization through univariate volume analysis were maintained using voxel-level harmonization: group difference in FA_{Low} at 3- and 30-days post injury were maintained ($p < 0.01$), as well as site-specific differences in FA_{High} at 3- and 30-days post injury ($p < 0.01$) (Figure 4).

Voxel-level harmonization improves group-level delineation of injured brain regions and leads to spatially specific statistical improvements

Since the application of NeuroCombat at the voxel level resulted in decreased cross-site differences, supporting prior work at the same site across different field strengths (17), we next investigated the statistical effect of harmonization within regions of primary injury as well as in less injured regions across the brain. Correct interpretation of statistical results requires consideration not only of statistical significance but effect size and statistical power. Improvements in power, the probability of detecting a result when it is present, can help to detect significant differences in datasets with small effect sizes. Given the simple statistical design of the current study, increases in sample size from combining data in this study should generally be expected to increase statistical power at a geometric rate through reduction in standard error.

However, power is also affected by effect size and unexplained variances, among other factors. Technical/hardware variation and site related effects when measuring the injury present themselves as ideal sources of variance for NeuroCombat to capture. We aimed to understand how cross-site harmonization was able to resolve this variance, as well as how harmonization affected effect size and power across the brain. We hypothesized that harmonization would increase the probability of correctly detecting group-level differences, particularly within specific brain regions that may exhibit high variability at the single site level through increases in power and effect size.

To accomplish this, whole brain images of change in power and effect size due to harmonization were generated by voxel-based subtraction between the original, non-harmonized and the harmonized FA data for each site (Figure 7). The average effect from harmonization calculated over all three sites was an increase in power around the primary injury area of the ipsilateral corpus callosum but decreases in most other regions (Figure 7A). At the site level at day 3, power increased within the ipsilateral white matter and gray matter surrounding the primary injury area in all three site data, with the greatest improvement measured in Site 1 (Figure 7B). There were, however, decreases in power as a result of harmonization, globally within the grey matter across all sites and at both times post-injury. At day 30, the greatest increase in power was observed in data from Site 1, similar to day 3, while there were power decreases in the striatum and ventral areas of the brain in data from Sites 2 and 3. These same remote changes also generalized to the AD, MD and RD (Supplementary Figures S4, S9, S14).

Calculation of the corresponding voxel-wise change in effect size revealed an ipsilateral increase in the magnitude in the site-averaged data (Figure 8A) consistent with the prior power changes (Figure 7A).

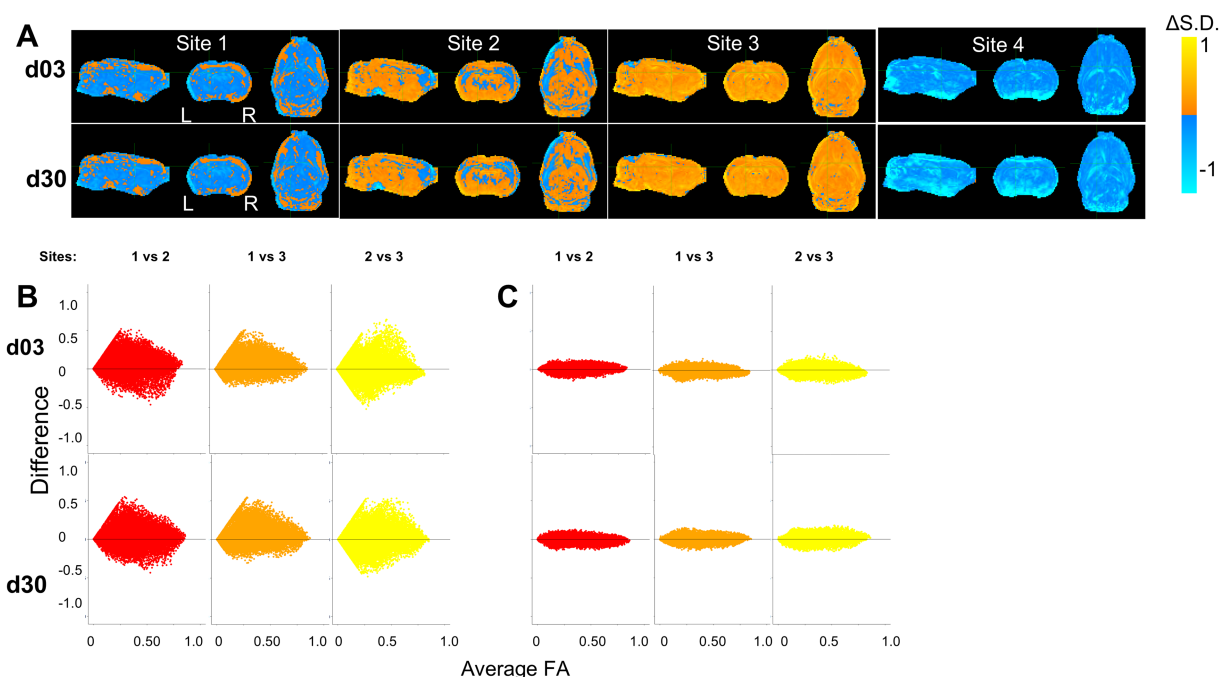


FIGURE 6

Harmonization at voxel-level resolution. **(A)** Change in standard deviation due to harmonization (harmonized – non-harmonized), where orange and blue voxels indicate an increase and decrease in standard deviations due to harmonization, respectively. Site 4 shows a decrease in standard deviation across the brain, indicating a large variation in the original, non-harmonized sham distribution of FA values across the brain. Key: [L] = Left side of the brain, [R] = Right side of the brain. **(B,C)** Harmonization of FA values of each individual voxel across site pairs is demonstrated using Bland–Altman plots for **(A)** original, unharmonized values and **(B)** harmonized values. The difference between sites decreases due to harmonization.

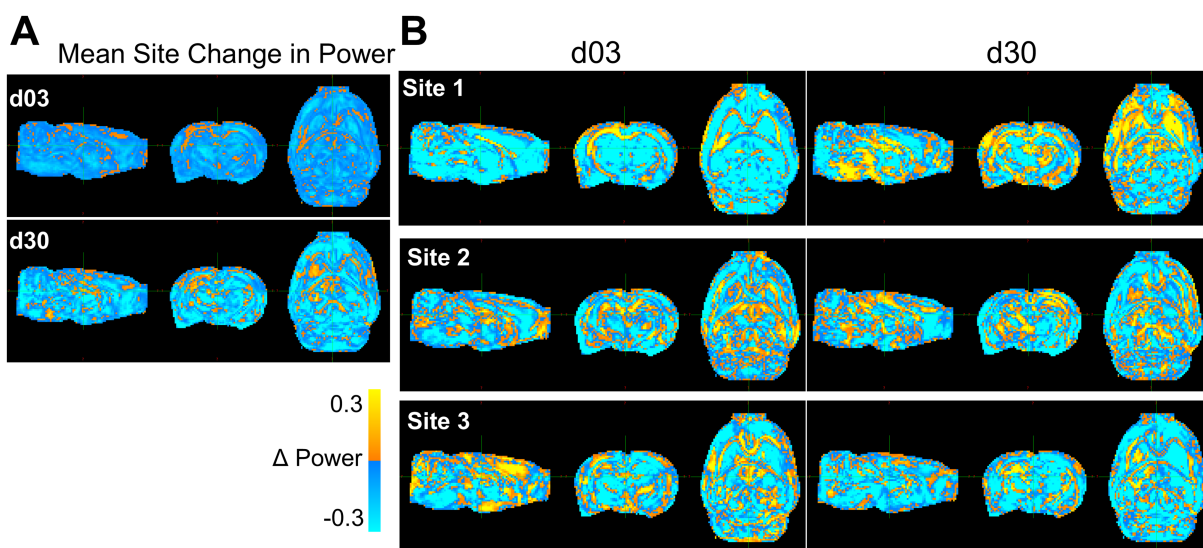


FIGURE 7

Whole-brain maps of regional changes in power due to harmonization. **(A)** Difference in power across a pooled average of all three sites post-injury due to harmonization and **(B)** across site and post-injury day. The magnitude scale –30 and 30% indicates the percent difference in statistical power between pre and post voxel-level harmonization, where yellow/blue indicate power increases and decreases, respectively due to harmonization.

However, while power changes were decreased by harmonization throughout much of the brain, effect size increased over the majority of the brain remote from the primary injury site, with the exception of decreases within multiple clustered voxels that occurred at both

post-injury times. Across site-averaged data, these trends persisted in AD, MD and RD ([Supplementary Figures S4, S9, S14](#)), but the drivers varied. For FA, the underlying data driving these site-wide effects were a decrease in effect size at site 1, but an increase at Sites 2 and 3

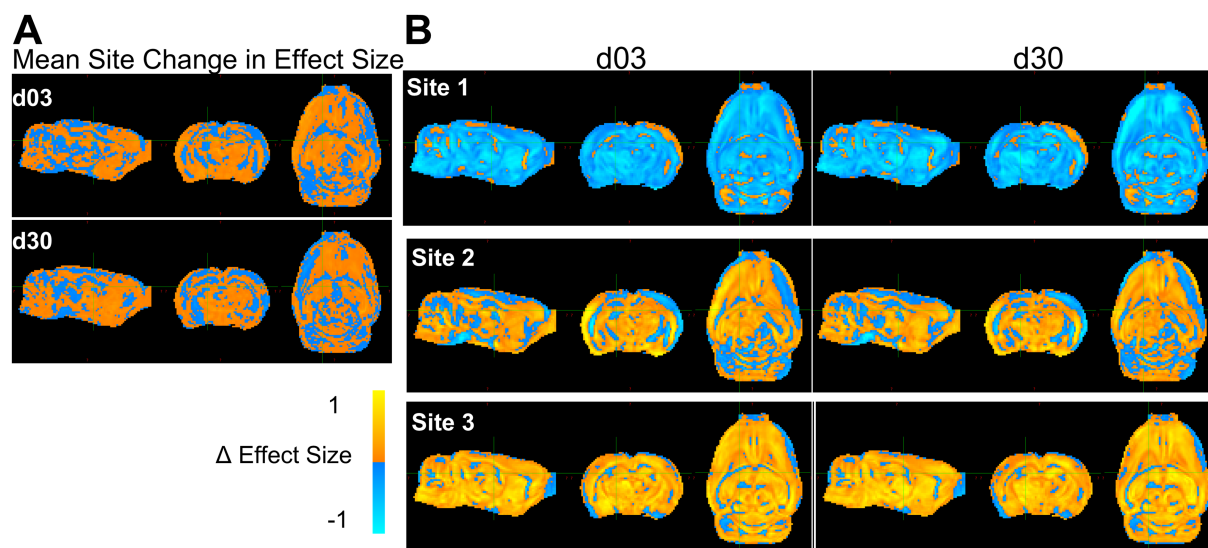


FIGURE 8

Whole-brain maps of regional changes in effect size due to harmonization. (A) Difference in effect size across a pooled average of all three sites post-injury due to harmonization and (B) across sites and post-injury day. The magnitude scale -1 and $+1$ Cohen's d indicates the difference in effect size/voxel between pre and post voxel-level harmonization, where yellow/blue indicate increases and decreases, respectively due to harmonization.

(Figure 8B). The opposite trend was seen in MD and RD, where Site 1 showed increases in effect size in the injury area, while Sites 2 and 3 exhibited large decreases around the injury area (Supplementary Figures S9, S14). AD showed a different trend, with Site 2 driving much of the decreases and Sites 1 and 3 showing increased effect size at the injury area (Supplementary Figure S4). The heterogeneity in both effect size and power across the sites outside of the primary injury area may be explained by the effect of NeuroCombat as it attempts to adjust for the variability introduced by site effects, ultimately converging at a common spatial pattern.

We sought to determine whether the improved power and effect size was underpinned by the detection of pathologically low or high FA when compared to the sham group. FA voxel overlap maps showing the proportion of rats with FA values significantly different from sham after harmonization versus before, revealed an increased number, and overlap of rats after harmonization when compared to the original data (Figure 9). These increases were especially prevalent within, and adjacent to, the primary, ipsilateral injury area. All differences detected were related to decreases in FA; no data was underpinned by increased areas of FA_{High} when compared to shams. These results indicate that cross-site data harmonization can improve the reliability of group differences in the areas surrounding the primary injury area at the cost of lower power in regions more remote from the primary injury site. For other scalar measures, results were markedly different. Generally, the proportion of rats significantly different from mean sham was lower. In the AD dataset, most areas of difference were found in the ipsilateral cortex, frontal pole and cerebellum. The only location to produce an increase in the number of rats was a small portion of the corpus callosum. In other areas, harmonization decreased the number of voxels with significant differences, suggesting an initially weak effect at both day 3 and 30 (Supplementary Figure S5). Similarly, the MD data at day 3 showed large portions of the cortex with a small number of rats being significantly different from sham, but this area decreased after harmonization. At day 30, the number of voxels increased around pre-existing areas of difference in the cerebellum and parts of the

ipsilateral cortex (Supplementary Figure S10). In the RD dataset, areas of pre-existing cortical differences decreased after harmonization at day 3 and day 30, except for minor increases in coverage in the contralateral cerebellum (Supplementary Figure S15).

Discussion

This study demonstrates that NeuroCombat harmonization improved the detection of injury after univariate-level harmonization, although this effect was greater using pooled rather than site-specific shams. Despite increases in effect size due to harmonization, group differences were both gained and lost dependent on scalar volume, directional difference compared to sham, and post-injury time. On the other hand, voxel-level harmonization retained group differences present within the original data and led to increases in statistical power and effect size to detect group differences within the primary injured area. These voxel locations displayed a higher frequency of decreased FA in injured rats, but at the potential cost of decreased reliability to detect group differences within regions remote from the injury site.

We found that NeuroCombat harmonization does decrease site-specific effect in univariate and voxel-wise measures of FA, thus better delineating biological differences as a result of injury. Analysis further displayed improved harmonization with the use of a pooled sham population and removal of significant outliers.

Site outliers severely bias NeuroCombat

A significant challenge to NeuroCombat harmonization is the presence of outliers. These findings are consistent with prior reports (43) that the presence of outliers can severely bias the ability of NeuroCombat to remove site effects. That prior work identified outliers as measurements outside 3 interquartile ranges (IQRs) of Q1 or Q3, and determined that if more than 5% of values at a particular

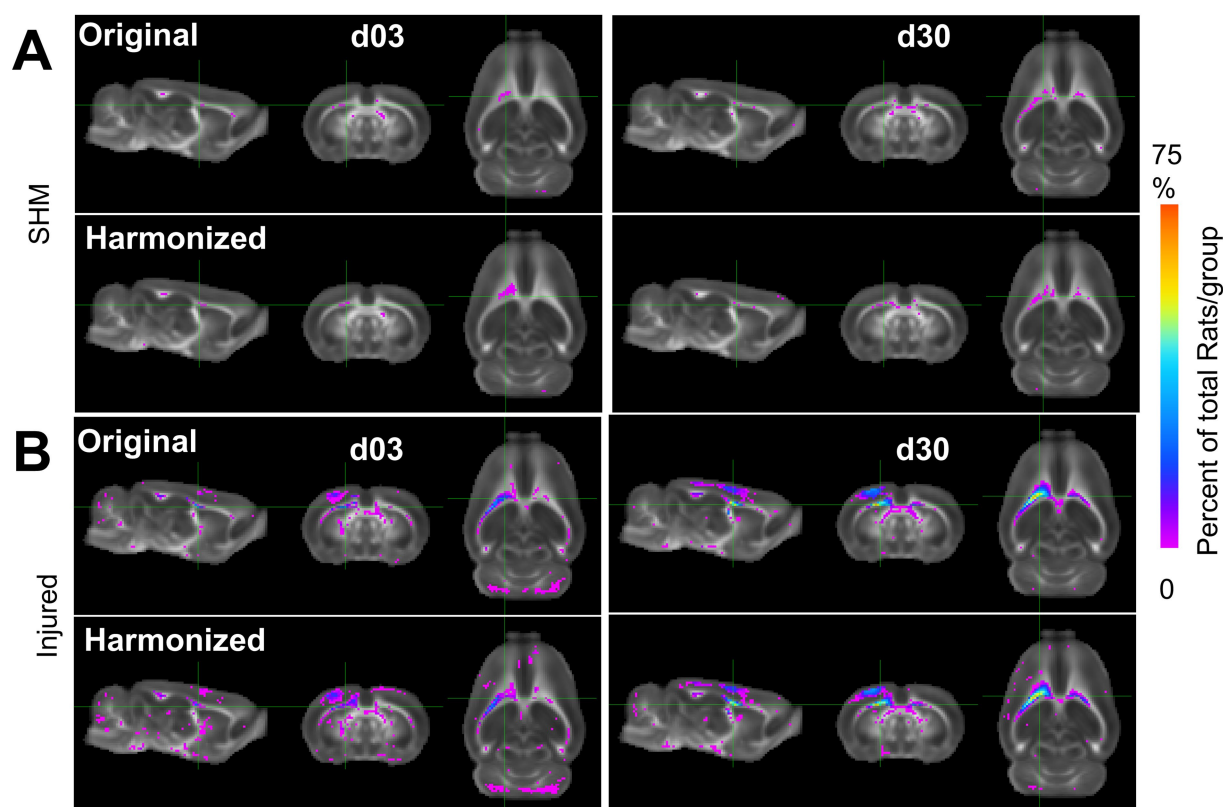


FIGURE 9

Multi-Site Voxel overlap maps delineating larger regions of common pathology due to data harmonization. (A) Sham and (B) Injured voxel overlap maps derived from whole-brain data with all three sites showing the incidence of the number of rats at each voxel location where FA values differ significantly from pooled sham data before (label-Original) and after harmonization (label-Harmonized). Images show voxels where there was a lower (Pink) and higher (Blue/Red) proportion of CCI rats in which the FA value was significantly different from pooled shams due to data harmonization ($p < 0.01$, FDR corrected).

site were outliers, then the ability for NeuroCombat to harmonize data was hindered, and this was associated with reduced variability within the outlier site data (43). Site 4 in the current study exhibited these characteristics with an outlier proportion of $>10\%$ when outliers were measured as 1.5 IQRs outside of Q1 and Q3. Difference in scanner field strength is unlikely to be the source of variation since another site that operated at the same field strength (Site 3) showed no such variation. One major difference in hardware between Site 4 and all other sites was the use of different radiofrequency coil hardware that resulted in a reduced brain coverage that may have accounted for the increased variation (Figure 2). A second major difference was the lower number of rats included from Site 4 in both experimental groups, making the statistical correction to the z-scored volume data larger and thus less sensitive to the detection of abnormal diffusion scalar-related values. This suggests that a more conservative approach during pre-acquisition harmonization is necessary when selecting if site data can be grouped together for harmonization. This may relate to both MRI hardware and injury severity level.

Choice of site-specific versus pooled shams is important

Use of pooled sham data from across all three sites, rather than a site-specific approach, produced a greater percentage of TBI rats

where the volume of pathology indicated by FA_{Low} or FA_{High} was greater than in sham rats. This was counter to expectations, where our initial hypothesis was that site-specific sham data would encode site-specific technical, scanner-related effects that would be negated through within-site identification of TBI volumes of pathology, leading to more uniform cross-site data. This premise was based on the current lack of available preclinical phantoms that could be used for correcting for site-specific effects, leaving us to make use of local sham data to fill this gap. However, despite the greater success of using pooled shams, it was only the use of site-specific shams that retained group differences in *univariate-level harmonized* FA_{Low} volumes. Although Combat-seq was unable to completely resolve site effects in some univariate data, it did maintain or enhance the strong group differences. Because a similar group effect was found through voxel harmonization, it may indicate that Combat-seq is able to improve injury detection within univariate data. This may suggest that it is more suitable compared to NeuroCombat for harmonizing univariate, count-distributed neuroimaging data. This is the first study that the authors are aware of that has applied Combat-seq to univariate statistics derived from neuroimaging data. Combat-seq was previously only applied to RNA sequencing data. Because the voxel-level harmonization data did possess site effects, harmonization was able to preserve true biological variation, which prevented the loss of signal. Despite this finding, *voxel-level harmonization* maintained all group differences present within the original data. It is possible that the lower

sample size at the single-site level could lead to effects that may either overestimate or underestimate true effects, where pooling data can increase sensitivity to true effects (44). It could also be that the whole-brain data used to generate voxel-level harmonization metrics was continuous, which enabled more successful harmonization with NeuroCombat. It therefore remains crucial to study the effect of how sham data are used when determining the effects of cross-site harmonization.

Harmonization improves power and effect size in areas with existing biological differences

Consistent with prior literature (12), voxel-level harmonization effectively removed site effects in whole-brain data. As shown in the Bland–Altman plots (Figures 6B,C), across all combinations of sites and mean FA value of voxel pairs, the difference between each pair of voxels was reduced after harmonization. This reduction in variability across sites after harmonization corresponds with results seen in both the preclinical and clinical spaces (12, 17, 45). These changes suggest that harmonization improves the ability to detect group differences in areas of pre-existing biological variation. In areas where no significant differences between sham and injury were initially found, power decreased. Univariate measures of pathology showed similar results, with mild effect size increases between injured and sham groups after harmonization. This could explain some of the detrimental effects of harmonization, which has been shown to obscure biological variation in data without site effects, such as in regions unaffected by injury or site (16). The current finding that the proportion of rats which were significantly different from shams did not increase substantially outside of brain regions where group differences were detected before harmonization suggests that NeuroCombat does not introduce erroneous biological variation (14, 15). Voxel-level harmonization appears to only be able to improve detection of group differences across data where group differences already exist, rather than revealing differences obscured by noise or systematic error.

In areas of the brain where group differences were already present, NeuroCombat successfully improved delineation of injury among FA datasets, but not MD, RD or AD. Improvements in power and effect size were seen in areas of the ipsilateral white matter and cortex, which generally corresponds to areas of existing group differences in the FA data. This was further seen by improvements in the proportion of rats that exhibited group differences at these voxels at day 3 and day 30. The use of NeuroCombat on data with site effects may help to improve reproducibility in this respect, as improved power, effect size and more frequent detection of injury can help to reduce erroneous conclusions. An increase in power indicates a benefit for pooling data across sites, allowing for smaller samples from each of multiple sites to support observed effects. The decrease in power observed in non-injured areas could have been related to a lack of biological effect in those areas prior to harmonization, where harmonization increased power in areas of existing difference at the expense of whole-brain differences. An alternative reason could be related to a decrease in inaccurate effect sizes within non-injured regions, which would cause power to decrease in those areas and reflect areas of true biological difference. Similar findings have been reported in clinical FA data: harmonization with NeuroCombat was able to increase the number

of voxels associated with age in a clinical multi-site and multi-scanner study of DTI measures' relationship with age (12), as well as improve power in the detection of differences among individuals with schizophrenia in T1 derived metrics of cortical thickness, surface area and subcortical volumes (5). The present study builds upon these initial findings by validating NeuroCombat's efficacy in preclinical data, as well as by demonstrating that these improvements in the detection of group differences are confined within areas of existing biological variation.

There was a relative lack of success by harmonization to improve injury detection among AD, MD and RD datasets compared to FA, despite varying degrees of power and effect size improvements. Reasons for this are not clear, but it may relate to site differences in measurement of absolute diffusivities compared to a more equitable comparison of FA which is a normalized ratio of these diffusivities, and thus less likely to differ in absolute values across site. Further examination into whether normalization of unbounded metrics helps to improve the efficacy of NeuroCombat may be warranted to better understand this phenomenon. Other methods of ComBat that have been used in imaging data include generalized additive models (gamCombat) which accounts for non-linear covariates, longCombat which can be used to detect changes over time, and gaussian mixture model-combat that removes variation due to unidentified covariates (16, 46). NeuroCombat has been shown to be preferable when conducting cross-sectional data harmonization compared to longCombat (16). While in the present study we were focused on group-level comparisons at each timepoint, the use of longCombat may well be useful to capture differences in injury trajectories across site in future work.

Limitations

Inherent in the preprocessing of image data are the potential errors associated with the requirement to conduct spatial transformations to ensure correct registration to a common 3D space. Multiple transformations could have impacted the quality and precision of the resulting measures of pathology. The heterogeneous nature of injury may also play a role in the decrease in power in non-significantly different brain areas. Due to potential minor differences in delivery of trauma, brain areas of some rats may have been more injured compared to similar areas in other rats. This may lead to less consistent measurements and interfere with NeuroCombat's abilities to resolve site effects.

Univariate level volume analysis was conducted after identifying regions of $\text{Scalar}_{\text{Low}}$ and $\text{Scalar}_{\text{High}}$ through a statistical-based z-corrected threshold (30, 31). However, there is no agreed upon statistical correction required for the analysis of voxel-based data. It is possible that the strict threshold used here ($p < 0.001$) reduced the ability to accurately judge pathologic tissue within injured versus sham rats, as indicated by the low overlap between rats in site 4 (Figure 2). More work is required to determine the impact of statistical correction threshold on determining group differences after TBI.

The biological basis for FA_{HIGH} regions after TBI is uncertain, although it has been found before in the same spatially discrete patterns from FA_{LOW} regions (3). Noteworthy here though is that the benefit provided by multi-site harmonization was afforded by FA_{LOW} , and from which the biological underpinnings is far more certain.

Conclusion

NeuroCombat harmonization demonstrates utility in reducing site effects in rodent imaging across multiple sites and scanner field strengths. By increasing the statistical power and effect size to detect areas of injury, post-acquisition, cross-site data harmonization improves the ability to discriminate between sham and injured rats. This should enable improved efficiency for preclinical study completion by collecting data at multiple sites.

Significance statement

This project demonstrates the utility of NeuroCombat in reducing site effects in multi-site rodent imaging. We also demonstrate that harmonization improves the ability to distinguish between sham and injured rats at the voxel level and increase statistical power and effect size in areas of injury. Multi-center studies are becoming more common to allow for increased efficiency in data collection, and with conservative approaches and analysis into the datasets, NeuroCombat can be utilized to improve study reliability and reproducibility.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. Univariate data derived in this study are available through the Open Data Commons for Traumatic Brain Injury (odc-tbi.org; RRID:SCR_021736), Kislik et al. (2025), [http://doi.org/10.34945/F5D60C](https://doi.org/10.34945/F5D60C).

Ethics statement

The animal study was approved by each site-specific animal research committee: UCLA approval #2021-082; UF/MSM approval # 202010145; GU/USU approval APG-21-044; JHU approval #RA22M370. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

GK: Methodology, Writing – review & editing, Writing – original draft, Visualization, Formal analysis. RF: Writing – review & editing, Methodology, Formal analysis, Visualization, Writing – original draft. AK: Data curation, Writing – review & editing. JZ: Supervision, Funding acquisition, Software, Writing – review & editing. MF: Data curation, Writing – review & editing. BM: Project administration, Data curation, Writing – review & editing. AB: Writing – review & editing, Data curation. YZ: Data curation, Writing – review & editing. JW: Writing – review & editing, Data curation. RK: Writing – review & editing, Funding acquisition. TA: Writing – original draft, Data curation, Writing – review & editing. MB: Writing – review & editing, Funding acquisition. JM: Writing – review & editing, Funding acquisition. KW: Funding acquisition, Writing – review & editing. JH: Funding acquisition, Writing – review & editing. AF: Writing – review

& editing, Funding acquisition. AP: Writing – review & editing, Data curation, Methodology. IW: Investigation, Writing – review & editing, Funding acquisition. NH: Visualization, Funding acquisition, Resources, Conceptualization, Writing – review & editing, Formal analysis, Project administration, Supervision, Writing – original draft, Methodology.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Supported by National Institute of Neurological Disorders and Stroke Translational Outcomes Project for Neurotrauma (TOP-NT) Awards NS106945 (to N.G.H. and I.B.W.), NS106941 (to M.P.B.), NS106937 (to J.Z. and RCK), NS106899 (to A.R.F.), and NS106938 (to K.K.W.) and the Vivian L. Smith Foundation.

Acknowledgments

The TOP-NT Investigators include (in alphabetical order): Albenese, Chris, Georgetown University, Washington, DC. Buenaventura, Ruchelle, Georgetown University, Washington, DC. Dickstein, Dara L., Uniformed Services University of the Health Sciences, Bethesda, MD. Fan, Anya, Uniformed Services University of the Health Sciences and the Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD. Fu, Amanda, Uniformed Services University of the Health Sciences, Bethesda, MD and the Henry M. Jackson Foundation for the Advancement of Military Medicine Harvey, Alexander C., Georgetown University, Washington, DC. Labastida, Javier Allende, Johns Hopkins University, Baltimore, Maryland. Lee, Yichien, Georgetown University, Washington, DC. Liu, Jiong, Uniformed Services University of the Health Sciences and the Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD. Klee, Madison P., University of California Los Angeles, Los Angeles, CA. Main, Bevan S., Georgetown University, Washington, DC. Rodruigez, Olga, Georgetown University, Washington, DC. Tong, Jonathan, University of California Los Angeles, Los Angeles, CA. Torres-Espín, Abel, University of Waterloo, Waterloo, ON.

Tucker, Laura B., Uniformed Services University of the Health Sciences and the Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD. Vanderveer-Harris, Nathan, University of California Los Angeles, Los Angeles, CA.

Conflict of interest

J. M is a federal employee, but the opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the U.S. Army, Department of Defense, the U.S. Government, the Uniformed Services University of the Health Sciences, or Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. The use of trade names does not constitute an official endorsement or approval of the use of reagents or commercial hardware or software. The content is solely the responsibility of the authors and does not necessarily represent the official views of the

National Institute of Neurological Disorders and Stroke, or the National Institutes of Health and the U.S. Department of Health and Human Services.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

References

1. Albeni BC, Knobloch SM, Chew BGM, O'Reilly MP, Faden AI, Pekar JJ. Diffusion and high resolution MRI of traumatic brain injury in rats: time course and correlation with histology. *Exp Neurol*. (2000) 162:61–72. doi: 10.1006/exnr.2000.7256
2. Benson RR, Meda SA, Vasudevan S, Kou Z, Govindarajan KA, Hanks RA, et al. Global white matter analysis of diffusion tensor images is predictive of injury severity in traumatic brain injury. *J Neurotrauma*. (2007) 24:446–59. doi: 10.1089/neu.2006.0153
3. Harris NG, Verley DR, Gutman BA, Sutton RL. Bi-directional changes in fractional anisotropy after experiment TBI: disorganization and reorganization? *Neuro Image*. (2016) 133:129–43. doi: 10.1016/j.neuroimage.2016.03.012
4. Basser PJ, Pierpaoli C. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J Magn Reson B*. (1996) 111:209–19. doi: 10.1006/jmrb.1996.0086
5. Radua J, Vieta E, Shinohara R, Kochunov P, Quidé Y, Green MJ, et al. Increased power by harmonizing structural MRI site differences with the com bat batch adjustment method in ENIGMA. *Neuro Image*. (2020) 218:116956. doi: 10.1016/j.neuroimage.2020.116956
6. Dansereau C, Benhajali Y, Risterucci C, Pich EM, Orban P, Arnold D, et al. Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. *Neuro Image*. (2017) 149:220–32. doi: 10.1016/j.neuroimage.2017.01.072
7. Radabaugh HL, Harris NG, Wanner IB, Burns MP, McCabe JT, Korotcov AV, et al. Translational outcomes project in neurotrauma (TOP-NT) pre-clinical consortium study: a synopsis. *J Neurotrauma*. (2025) 42:898–912. doi: 10.1089/neu.2023.0654
8. Kamnakh A, Puhakka N, Ali I, Smith G, Aniceto R, McCullough J, et al. Harmonization of pipeline for preclinical multicenter plasma protein and miRNA biomarker discovery in a rat model of post-traumatic epileptogenesis. *Epilepsy Res*. (2019) 149:92–101. doi: 10.1016/j.eplepsyres.2018.11.009
9. Magnotta VA, Matsui JT, Liu D, Johnson HJ, Long JD, Bolster BD, et al. Multicenter reliability of diffusion tensor imaging. *Brain Connect*. (2012) 2:345–55. doi: 10.1089/brain.2012.0112
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. (2016) 3:160018. doi: 10.1038/s41597-019-0009-6
11. Hu F, Chen AA, Horng H, Bashyam V, Davatzikos C, Alexander-Bloch A, et al. Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *Neuro Image*. (2023) 274:120125. doi: 10.1016/j.neuroimage.2023.120125
12. Fortin JP, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*. (2017) 161:149–70. doi: 10.1016/j.neuroimage.2017.08.047
13. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. (2007) 8:118–27. doi: 10.1093/biostatistics/kjx037
14. Wengler K, Cassidy C, van der Pluijm M, Weinstein JJ, Abi-Dargham A, van de Giessen E, et al. Cross-scanner harmonization of Neuromelanin-sensitive MRI for multisite studies. *J Magn Reson Imaging*. (2021) 54:1189–99. doi: 10.1002/jmri.27679
15. Bell TK, Godfrey KJ, Ware AL, Yeates KO, Harris AD. Harmonization of multi-site MRS data with ComBat. *NeuroImage*. (2022) 257:119330. doi: 10.1016/j.neuroimage.2022.119330
16. Richter S, Winzeck S, Correia MM, Kornaropoulos EN, Manktelow A, Outtrim J, et al. Validation of cross-sectional and longitudinal ComBat harmonization methods for

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fneur.2025.1612598/full#supplementary-material>

- magnetic resonance imaging data on a travelling subject cohort. *Neuroimage Rep*. (2022) 2 None. doi: 10.1016/j.jnirp.2022.100136
17. Quach M, Ali I, Shultz SR, Casillas-Espinosa PM, Hudson MR, Jones NC, et al. Combating inter-site differences in field strength: harmonizing preclinical traumatic brain injury MRI data. *NMR Biomed*. (2024) 37:e5142. doi: 10.1002/nbm.5142
18. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*. (2020) 2:lqaa078. doi: 10.1093/nargab/lqaa078
19. Wanner IB, McCabe JT, Huie JR, Harris NG, Paydar A, McMann-Chapman C, et al. Prospective harmonization, common data elements, and sharing strategies for multicenter pre-clinical traumatic brain injury research in the translational outcomes project in Neurotrauma consortium. *J Neurotrauma*. (2025) 42:877–897. doi: 10.1089/neu.2023.0653
20. du Sert NP, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *PLoS Biol*. (2020) 18:e3000410. doi: 10.5281/zenodo.3877179
21. Lee SH, Ban W, Shih YYI. BrkRaw/bruker: BrkRaw v0.3.3 [Internet]. Zenodo; (2020) (cited 2025 Mar 31). Available online at: <https://zenodo.org/records/3877179>
22. Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*. (2002) 17:479–89. doi: 10.1006/nimg.2002.1040
23. Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*. (2014) 90:449–68. doi: 10.1016/j.neuroimage.2013.11.046
24. Kellner E, Dhital B, Kiselev VG, Reiser M. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magn Reson Med*. (2016) 76:1574–81. doi: 10.1002/mrm.26054
25. Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*. (2003) 20:870–88. doi: 10.1016/S1053-8119(03)00336-7
26. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*. (2004) 23:S208–19. doi: 10.1016/j.neuroimage.2004.07.051
27. Woolrich MW, Jbabdi S, Patenaude B, Chappell M, Makni S, Behrens T, et al. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*. (2009) 45:S173–86. doi: 10.1016/j.neuroimage.2008.10.055
28. Tournier JD, Smith R, Raffelt D, Tabbara R, Dhollander T, Pietsch M, et al. MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*. (2019) 202:116137. doi: 10.1016/j.neuroimage.2019.116137
29. 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) 54:2033–44. doi: 10.1016/j.neuroimage.2010.09.025
30. Mayer AR, Bedrick EJ, Ling JM, Toulouse T, Dodd A. Methods for identifying subject-specific abnormalities in neuroimaging data. *Hum Brain Mapp*. (2014) 35:5457–70. doi: 10.1002/hbm.22563
31. Smith G, Santana-Gomez CE, Staba R, Harris NG. Unbiased population-based statistics to obtain pathologic burden of injury after experimental TBI. *Exp Neurol*. (2025) 392:115332. doi: 10.1016/j.expneurol.2025.115332

32. Verley DR, Torolira D, Pulido B, Gutman B, Bragin A, Mayer A, et al. Remote changes in cortical excitability after experimental traumatic brain injury and functional reorganization. *J Neurotrauma*. (2018) 35:2448–61. doi: 10.1089/neu.2017.5536
33. Paydar A, Harris NG. The pericontused cortex can support function early after TBI but it remains functionally isolated from normal afferent input. *Exp Neurol*. (2023) 359:114260. doi: 10.1016/j.expneurol.2022.114260
34. Avants BB, Tustison NJ, Stauffer M, Song G, Wu B, Gee JC. The insight tool kit image registration framework. *Front Neuroinform*. (2014) 8:44. doi: 10.3389/fninf.2014.00044
35. Rane Levendovszky S. Cross-sectional and longitudinal hippocampal atrophy, not cortical thinning, occurs in amyloid-negative, p-tau-positive, older adults with non-amyloid pathology and mild cognitive impairment. *Front Neuroimag*. (2022) 1:828767. doi: 10.3389/fnimg.2022.828767
36. Maikusa N, Zhu Y, Uematsu A, Yamashita A, Saotome K, Okada N, et al. Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. *Hum Brain Mapp*. (2021) 42:5278–87. doi: 10.1002/hbm.25615
37. Wickham H. *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer (2009).
38. Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: data analysis with estimation graphics. *Nat Methods*. (2019) 16:565–6. doi: 10.1038/s41592-019-0470-3
39. Whitcher B, Schmid VJ, Thornton A. Working with the DICOM and NIFTI data standards in R. *J Stat Softw*. (2011) 44:1–29.
40. Muschelli J. Neurobase: “Neuroconductor” base package with helper functions for “nifti” objects. (2024) [cited 2025 Mar 31]. Available online at: <https://cran.r-project.org/web/packages/neurobase/>
41. McCarthy P. FSLeys [Internet]. Zenodo; (2024) [cited 2025 Mar 30]. Available from: <https://zenodo.org/records/11047709>
42. Champely S. (2006). Pwr: Basic functions for power analysis [internet]. [cited 2025 Mar 30]. p. 1.3–0. Available online at: <https://CRAN.R-project.org/package=pwr>
43. Han Q, Xiao X, Wang S, Qin W, Yu C, Liang M. Characterization of the effects of outliers on ComBat harmonization for removing inter-site data heterogeneity in multisite neuroimaging studies. *Front Neurosci*. (2023) 17:1146175. doi: 10.3389/fnins.2023.1146175
44. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. (2013) 14:365–76. doi: 10.1038/nrn3475
45. Yu M, Linn KA, Cook PA, Phillips ML, McInnis M, Fava M, et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum Brain Mapp*. (2018) 39:4213–27. doi: 10.1002/hbm.24241
46. Horng H, Singh A, Yousefi B, Cohen EA, Haghighi B, Katz S, et al. Improved generalized ComBat methods for harmonization of radiomic features. *Sci Rep*. (2022) 12:19009. doi: 10.1038/s41598-022-23328-0