The Anatomical Tracings of Lesions After Stroke (ATLAS) Dataset - Release 1.1

Authors

Sook-Lei Liew[1†*], Julia M. Anglin[1*], Nick W. Banks[1], Matt Sondag[1], Kaori L. Ito[1], Hosung Kim[1], Jennifer Chan[1], Joyce Ito[1], Connie Jung[1], Stephanie Lefebvre[1], William Nakamura[1], David Saldana[1], Allie Schmiesing[1], Cathy Tran[1], Danny Vo[1], Tyler Ard[1], Panthea Heydari[1], Bokkyu Kim[1], Lisa Aziz-Zadeh[1], Steven C. Cramer[2], Jingchun Liu[3], Surjo Soekadar[4], Jan-Egil Nordvik[5], Lars T. Westlye[6,7], Junping Wang[3], Carolee Winstein[1], Chunshui Yu[3], Lei Ai[8], Bonhwang Koo [8], R. Cameron Craddock[8,9], Michael Milham[8,9], Matthew Lakich[10], Amy Pienta[11], Alison Stroud[11]

Affiliations

1. University of Southern California, Los Angeles, California, USA
2. University of California, Irvine, Irvine, California, USA
3. Tianjin Medical University General Hospital, Tianjin, China
4. University of Tübingen, Tübingen, Germany
5. Sunnaas Rehabilitation Hospital HT, Nesodden, Norway
6. NORMENT and KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway
7. Department of Psychology, University of Oslo, Oslo, Norway
8. Child Mind Institute, New York, New York, USA
9. Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New Jersey, USA
10. University of Texas Medical Branch, Galveston, Texas, USA
11. University of Michigan, Ann Arbor, Michigan

† Corresponding author: Sook-Lei Liew (sliew@usc.edu)

* Denotes equal contributions

Abstract

Stroke is the leading cause of adult disability worldwide, with up to two-thirds of individuals experiencing long-term disabilities. Large-scale neuroimaging studies have shown promise in identifying robust biomarkers (e.g., measures of brain structure) of stroke recovery. However, analyzing large datasets is problematic due to barriers in accurate stroke lesion segmentation. Manually-traced lesions are currently the gold standard for lesion segmentation, but are labor intensive and require anatomical expertise. While algorithms have been developed to automate this process, the results often lack accuracy. Newer algorithms that employ machine-learning techniques are promising, yet these require large training datasets to optimize performance. Here we present ATLAS (Anatomical Tracings of Lesions After Stroke), an open-source dataset of 304 T1-weighted MRIs with manually segmented lesions and metadata. This large, diverse dataset can be used to train and test lesion segmentation algorithms and provides a standardized dataset for comparing the performance of different segmentation methods. We hope ATLAS R1.1 will be a useful resource to assess and improve the accuracy of current lesion segmentation methods.

Background & Summary

Stroke is one of the leading causes of adult disability worldwide with up to two-thirds of stroke survivors experiencing long-term disabilities[1,2]. Despite intensive efforts to study stroke across a broad range of scientific disciplines, the recovery process after stroke is still not fully understood. Studying the brain and behavior of post-stroke populations may provide insights into this area. In particular, brain imaging provides a promising source of biomarkers (e.g., measures of brain structure or function) that could potentially predict an individual's likelihood of recovery and, more importantly, which treatments may be maximally effective for each individual[3–5]. This would allow for improved efficiency of resource utilization in clinical practice and clinical trials. Measures that include the size, location, and overlap of the lesion with existing brain regions or structures have been successfully used as predictors of stroke recovery[5–11]. However, to date, this has only been done in smaller-scale studies, and results may conflict across studies or be limited to each sample. Examining lesion properties with larger datasets could lead to the identification of more robust biomarkers that are widely applicable across diverse populations. Large-scale stroke neuroimaging datasets offer a promising approach to achieve a better understanding of the recovery process.

To properly analyze large-scale stroke neuroimaging datasets, however, accurate lesion segmentations for each individual are needed. Currently, the gold standard for identifying lesions is manual segmentation, a process that requires skilled tracers and can be prohibitively time consuming and subjective[12]. For example, a single large or complex lesion can up to several hours for even a skilled tracer. As a result of this demand on time and effort, this method, which has been used in previous smaller neuroimaging studies, is not suitable for larger sample sizes. Based on the literature, most studies with manually segmented brain lesions use smaller sample sizes between 10 to just over 100 brains[12–16]. Accurately segmenting hundreds of stroke lesions may thus present a barrier for larger-scale stroke neuroimaging studies.

Many stroke neuroimaging studies have utilized semi- or fully-automated lesion segmentation tools for their analyses. Semi-automated segmentation tools employ a combination of automated algorithms, which detect abnormalities in the MRI image, and manual corrections or inputs by an expert. Fully-automated algorithms rely completely on the algorithm for the lesion segmentation. Many of these fully-automated algorithms employ machine learning techniques that require training and testing on large datasets[17], and the performance of the algorithm is highly dependent on the size and diversity of the training dataset. However, there are few, if any, publically available large training/test datasets of manually segmented stroke lesion masks that could be used for improving such algorithms. Thus, while both semi- and fully-automated lesion segmentation tools have the potential to greatly reduce the time and expertise needed to analyze stroke MRI data[18], they currently lack the accuracy needed for rigorous stroke lesion-based analyses.

In addition, it is often hard to compare the performance of automated lesion segmentation tools as they are often not evaluated for performance on the same dataset. Recently, some exciting initiatives have emerged to

develop better segmentation algorithms using standardized datasets and metrics. In particular, the Ischemic Stroke Lesion Segmentation (ISLES) challenge is an annual satellite challenge of the Medical Image Computing and Computer Assisted Intervention (MICCAI) meeting that provides a standardized multimodal clinical MRI dataset of approximately 50-100 brains with manually segmented lesions[19]. The ISLES competition encourages research groups to use the dataset to evaluate their lesion segmentation algorithms and compare their results to that of other groups. This approach is promising for developing better lesion segmentation algorithms. However, ISLES challenge datasets have traditionally focused more on using multimodal clinical MRIs to predict more acute results. These algorithms are not easily translatable to the high-quality T1-weighted MRIs typically found in stroke rehabilitation research. Thus, here, we aimed to develop a complementary large dataset using only anatomical T1-weighted MRIs, which are typically acquired during research studies to assess rehabilitation outcomes. We anticipate this dataset could be useful for enhancing lesion segmentation methods for T1-weighted images often used medical rehabilitation research.

Here, we present ATLAS (Anatomical Tracings of Lesions After Stroke) Release 1.1, an open-source dataset consisting of 304 T1-weighted MRIs with manually segmented diverse lesions and metadata. The goal of ATLAS is to provide the research community with a standardized training and testing dataset for lesion segmentation algorithms on T1-weighted MRIs. This dataset can also be used to compare the performance of different lesion segmentation techniques. We believe that this diverse set of manually segmented lesions will serve as a valuable resource for researchers to use in assessing and improving the accuracy of lesion segmentation tools.

Methods

Overview

304 MRI images from 11 cohorts worldwide were collected from research groups in the ENIGMA Stroke Recovery Working Group consortium. Images consisted of T1-weighted anatomical MRIs of individuals after stroke. For each MRI, brain lesions were identified and masks were manually drawn on each individual brain in native space using MRIcron, an open-source tool for brain imaging visualization and defining volumes of interest (http://people.cas.sc.edu/rorden/mricron/index.html). A minimum of one lesion mask was identified for each individual MRI. If additional, separate (non-contiguous) lesions were identified, they were traced as separate masks. An expert neuroradiologist reviewed all lesions to provide additional qualitative descriptions of the type of stroke, primary lesion location, vascular territory, and intensity of white matter disease. Finally, a separate tracer performed quality control on each lesion mask. This included assessing the accuracy of the lesion segmentations, revising the lesion mask if needed, and categorizing the lesions to generate additional data such as the number of lesions in left and right hemispheres, and in cortical and subcortical regions. This dataset is provided in native subject space, with a subset of this dataset provided in standard space (normalized to the MNI-152 template).

Training Individuals Performing Lesion Tracing

Eleven individuals were trained in identifying and segmenting lesions. Training consisted of a detailed protocol and instructional video, and all tracers were guided through the training process by an expert tracer. All individuals were trained on an initial set of 5 brains with varying lesion sizes and locations (size range: min: 1,871 mm^3, max: 16,2015 mm^3; location: cortical [Left: 0, Right: 1], subcortical [Left: 4, Right: 0]). After tracing the first set of 5 lesions, tracings were reviewed by an expert tracer. One week later, individuals retraced the lesions on the same set of 5 brains, but were blinded to their first segmentation attempt to examine intra-tracer reliability. After this, each segmentation was reviewed and an expert tracer provided additional feedback. Inter- and intra-rater reliability measures and additional technical validation of the lesion tracings can be found in *Technical Validation* below.

Identifying and Tracing Lesions

To identify lesions, each T1-weighted MRI image was displayed using the multiple view option in MRIcron[20], which displays the brain in the coronal, sagittal, and axial view. To identify lesions, tracers looked for darker intensities within typically healthy tissue. For lesions that were more difficult to detect with the grayscale setting, colored look-up table settings (e.g., “cardiac”, “NIH”, or “spectrum” settings in MRIcron) were used to provide additional insight. Once the lesion or lesions were identified, the lesion mask was traced using either the coronal or axial view, using either a mouse, track pad, or tablet (i.e. Wacom Intuos Draw). A combination of MRIcron tools was used to draw the lesion masks, which included the 3D fill tool, the pen tool and the closed pen tool. Typically, and especially for larger sized lesions, tracers used the 3D fill tool to begin the segmentation. Crosshairs were placed in the center of the identified lesion and the tool would fill in voxels similar to the one at the point of origin with the selected radius and at the sensitivity specified by the difference from origin and difference at edge tools. The pen and closed pen tool, typically was used to fill in (or remove) the areas that the 3D tool had missed or was used to trace smaller lesions slice by slice. Once completed, lesion masks were saved in the volume of interest (VOI) file format with the identifier name “cXXXXsXXXXtXX_LesionRaw” (see *Data Records* below for full naming conventions). Lesions masks were then checked for correctness by a separate tracer and changes were made as needed. After lesions were identified as being correct, masks were smoothed using MRIcron’s smooth VOI tool where the full width half maximum (FWHM) parameter was set to 2 mm and the threshold was set to 0.5. These masks were saved in both the VOI (.voi) and NIFTI (.nii.gz) file format with the identifier name “cXXXXsXXXXtXX_LesionSmooth”. A probabilistic spatial overlap map showing the distribution of primary lesions can be found in Figure 1, and a 3D visualization of the map can be viewed in https://www.youtube.com/watch?v=Ag5CUsRNY9Q.

Any additional lesions that were not contiguous with the primary lesion mask were drawn as separate lesion masks and labeled. As described in *Data Records*, any secondary lesions followed the same procedures as the primary lesion mask, but were labeled as Lesion_1, Lesion_2, Lesion_3 and so on, with the naming convention moving from the largest to smallest mask (e.g., Lesion_1 is the largest secondary lesion mask). In general, the primary lesion mask was the largest lesion, with any secondary lesion masks subsequently named and ordered by size (largest to smallest). The only exception to this

was, in the case of multiple lesions, if the neuroradiologist identified a primary stroke location as a different lesion from the largest lesion mask. In these cases, we used the lesion identified by the neuroradiologist as the primary mask. This occurred in less than 5% of the subjects.

Metadata

For each lesion, we also provided metadata on the lesion properties to give the user additional qualitative information, beyond the binary lesion mask. This information can be used to quickly sort the dataset based on specific lesion characteristics (e.g., only left hemisphere lesions, or only subcortical lesions). It can also provide additional insight into the types of lesions that succeed or fail for a given lesion segmentation algorithm. The lesion properties were manually reported for each individual lesion mask. These include the number of lesions identified and traced, and the location of each lesion (i.e. right/left, subcortical, cortical, or other). In order to count each lesion only once, we defined subcortical lesions as lesions that are contained completely in the white matter and subcortical structures. Any lesion that extends beyond this area and into the cortex is considered a cortical lesion. In this way, cortical lesions may extend into the subcortical space, but subcortical lesions do not extend into the cortical space. “Other” includes the brainstem and cerebellum. An experienced neuroradiologist also identified the following information for each individual brain: the type of stroke (e.g., embolic, hemorrhagic), primary stroke location, vascular territory, and intensity of white matter disease (periventricular hyperintensities (PVH) and deep white matter hyperintensities (DWMH)). White matter hyperintensities were graded using the Fazekas scale[21]. For PVH, the following grades were applied: 0 = absence, 1 = “caps” or pencil-thin lining, 2 = smooth “halo”, 3 = irregular PVH extending into the deep white matter. For DWMH, the following grades were applied: a = absence, 1 = punctate foci, 2 = beginning confluence of foci, 3 = large confluent areas. The white matter hyperintensity ratings are included because areas of white matter hyperintensity often pose challenges for lesion segmentation algorithms.

Normalization to a Standard Template

Lesion segmentation algorithms vary in whether the input should be in native (subject) space or a standardized space. Therefore, to provide this option for users, we also generated a version of the ATLAS dataset in standard space. To convert the images to standard space, MRI images first underwent automated correction for intensity non-uniformity and intensity standardization using custom scripts derived from the MINC-toolkit[22] (https://github.com/BIC-MNI/minc-toolkit). These corrected images were linearly registered to the MNI-152 template using a version that was nonlinearly constructed and symmetric (version 2009; http://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009) to normalize their intracranial volume in a standardized stereotaxic space[23]. Using the resulting transformation matrix, the labels drawn on the MRI images were also registered to the MNI template. The MRI images were resampled using the linear interpolation whereas their labels used a nearest neighborhood interpolation to keep their binary nature. Due to technical difficulties, a subset of brains is not included in the standard space conversion, resulting in a total of n=239 ATLAS brains converted into standard MNI space. The standardized

ATLAS brains and masks are provided in a separate folder from the native space ATLAS brains and masks. A list of all the filename and naming conventions, along with file descriptors, can be found in Table 1.

Probabilistic Spatial Mapping of Lesion Labels

We also created a probabilistic spatial mapping of the lesion labels to visualize the distribution of lesion masks across the ATLAS dataset. To do this, we performed a population-based averaging of all the individual labels in MNI space, producing a voxel-wise map where values can range from 0 at each voxel (always background for all subjects) to 1 (100% presence of the lesion label across subjects). A probabilistic spatial map of the primary lesions can be found in Figure 1 and a 3D visualization of the lesion map can be found in the following video link: https://www.youtube.com/watch?v=Ag5CUsRNY9Q.

Defaced Dataset

Finally, to expand access to the dataset, we also created a defaced version of the standardized dataset in MNI space (n=229), which can be more widely accessed from the FCP-INDI archive. We used Freesurfer's mri_deface tools to perform the defacing (https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface) on all T1-weighted images. All images were named in accordance with the INDI data policy, and a meta-data sheet that links the INDI names of the defaced images to the standard ATLAS R1.1 dataset is included.

Data Records

The full dataset (*native* and *standard_mni* sets) is archived with the Archive of Disability Data to Enable Policy research (ADDEP) at the Inter-university Consortium for Political and Social Research (ICPSR). ICPSR is the world's largest social science data archive that supports several substantive-area archive collections including disability and rehabilitation. ICPSR provides access to the data and provides technical assistance to individuals accessing the data. In addition, a standardized, defaced subset of the dataset is archived with the International Data Sharing Initiative (INDI), which hosts many widely available neuroimaging datasets such as the Functional Connectome Project (FCP-INDI). See *Usage Notes* for more details regarding access.

For the full dataset archived with ICPSR, the naming convention and description of the files in ATLAS R1.1 can be found in Table 1. Within the ATLAS R1.1 main folder, there is an excel file with the metadata for the entire dataset. There are also two folders, one for the native space (subject space) MRIs (n=304), entitled *native,* and one for the standardized MRIs (n=239), entitled *standard_mni*. Throughout the dataset, MRIs are named and sorted based on each cohort (c); each cohort is in the format of cXXXX where XXXX is the number that the cohort was assigned (e.g., c0001). There are 11 total cohorts. Within each cohort folder are the individual subject (s) folders. Subject folders are named based on the cohort that they are in (cXXXX), the subject number that they were assigned (sXXXX) and the time point at which they were taken (tXX) (e.g. c0001s0004t01). For instance, participants with data taken two weeks apart would have two time points, where t01 is the first time point and t02 is the second.

Within the *native* folder, each subject folder has several components: at minimum, each will have the original T1-weighted MRI image (cXXXXsXXXXtXX.nii.gz) and three masks for the main lesion: the

unsmoothed lesion mask (cXXXXsXXXXtXX_LesionRaw.voi), and two smoothed lesion masks in (cXXXXsXXXXtXX_LesionSmooth.voi; cXXXXsXXXXtXX_LesionSmooth.nii.gz). The LesionRaw volume is the original hand-traced lesion volume, while the LesionSmooth volume used a Gaussian smoothing kernel (full width half maximum (FWHM) parameter set to 2 mm, threshold set to 0.5; see Methods above). We anticipated that most researchers would use the LesionSmooth volume as it is slightly more robust to small slice-by-slice human errors, and therefore created the .nii.giz version from this. Notably, the .voi files are in an MRIcron format so the masks can be further edited In MRICron if desired. The .nii.gz files use the standard NIfTI format[24] (http://nifti.nimh.nih.gov/nifti-1/), which can be opened, edited, and viewed by most standard neuroimaging software.

If a particular subject had multiple lesions, for each additional lesion, there would be three additional lesion masks (e.g. cXXXXsXXXXtXX_LesionRaw_1.voi, cXXXXsXXXXtXX_LesionSmooth_1.voi, cXXXXsXXXXtXX_LesionSmooth_1.nii.gz). In general, lesions were ranked based on size where the largest lesion was considered the main lesion. As mentioned previously, if the largest lesion differed from the primary lesion identified by the neuroradiologist, we deemed the primary lesion to be the one identified by the neuroradiologist. This occurred in less than 5% of cases.

The *standard_mni* folder follows the same file structure as the *native* folder. Here, each subject folder has several components, all in NIfTI format, and all normalized to MNI-152 template space. At a minimum, each subject folder has the standardized T1-weighted MRI image, registered to MNI-152 template space (cXXXXsXXXXtXX_t1w_stx.nii.gz) and a primary lesion mask, also registered to MNI-152 template space (cXXXXsXXXXtXX_LesionSmooth_stx.nii.gz). Additional lesion masks for that subject will have the same naming convention as above, with the appendix of "*stx.nii.gz" (e.g., cXXXXsXXXXtXX_LesionSmooth_2_stx.nii.gz).

Finally, in the FCP-INDI archive, there is the same file structure as the *standard_mni* folder from the ICPSR archive. The only difference here is that T1-weighted images are additionally defaced (n=229). There is also a separate naming convention, following the Brain Imaging Data Structure (http://bids.neuroimaging.io/), adopted by INDI and linked to the ATLAS naming convention.

Table 1 provides a list of all naming conventions and filenames, along with descriptions. Tables 2 and 3 provide a brief glance at the types of lesions found in the ATLAS dataset.

Table 1. Filenames and file descriptions for ATLAS R1.1 dataset. * represents a wildcard.

ICPSR ARCHIVE ***http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36684***	
***native* folder (n=304)**	
Filename or Identifier	**Description**
cXXXXsXXXXtXX.nii.gz	Raw T1-weighted MRI for each subject, where c = cohort number, s = subject number, and t = time point

*LesionRaw.voi	Raw primary lesion mask, drawn as a volume of interest in MRICron
*LesionSmooth.voi	Smoothed primary lesion mask, drawn as a volume of interest in MRICron
*LesionSmooth.nii.gz	Smoothed primary lesion mask, saved as a nifti file
*LesionRaw/Smooth_1(or 2, 3, …).voi/.nii.gz	Raw and smoothed secondary lesion masks (same as the three above, but for additional lesions)
***standard_mni* folder (n=239)**	
Filename or Identifier	**Description**
*t1w_stx.nii.gz	T1-weighted MRI for each subject, registered into MNI template space
*LesionSmooth_stx.nii.gz	Primary lesion mask, registered into MNI template space
*LesionSmooth_1(or 2, 3, ...)_stx.nii.gz	Secondary lesion masks, registered into MNI template space
standard_mni152.nii.gz	The MNI-152 template image used for registration
INDI ARCHIVE ***http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html***	
***standard* folder (n=229)**	
Filename or Identifier	**Description**
Site, Subject ID, Session	Naming convention follows Brain Imaging Data Structure (BIDS) recommendations; linked to ATLAS naming convention and meta-data described above

Table 2. There are a total of 304 subjects within 11 cohorts included in the full ATLAS Release 1.1 native dataset. The number of brains in which only one lesion was found (left/right hemispheres and other locations found within the brainstem and cerebellum, etc.), and the number of brains in which multiple lesions were found, are shown.

Cohort	Number of Subjects	Brains with One Lesion			Brains with Multiple Lesions
		Left	Right	Other	
c0001	6	3	2	0	1
c0002	25	6	12	1	6
c0003	55	14	19	0	22
c0004	34	7	3	1	23
c0005	30	9	3	6	12
c0006	12	4	1	2	5
c0007	36	9	9	0	18
c0008	32	8	11	0	13
c0009	12	8	0	0	4
c0010	47	9	10	7	21
c0011	15	1	11	0	3
Total	**304**	**78**	**81**	**17**	**128**

Table 3. The number of lesions found in each location (i.e. cortical vs. subcortical; left vs. right hemispheres), and other locations found within the brainstem and cerebellum, etc.) are shown. Here we have included primary lesions as well as additional lesions.

Cohort	Cortical Lesions		Subcortical Only Lesions		Other Lesion Locations
	Left	Right	Left	Right	
c0001	2	2	2	1	0
c0002	1	7	14	11	4
c0003	3	0	38	44	0
c0004	7	6	31	32	7
c0005	9	1	20	14	7
c0006	3	2	8	3	3
c0007	8	10	28	18	4
c0008	6	13	20	13	1
c0009	5	3	12	2	0
c0010	11	8	21	24	13
c0011	0	5	3	10	1
Total	**55**	**57**	**197**	**172**	**40**

Technical Validation

Each trained tracer created lesion masks for the same five brains twice, one week apart, to assess both inter- and intra-tracer reliability. Training lesions ranged in size and difficulty (see *Methods*). Each tracer's lesion masks were compared, providing both inter- and intra-rater reproducibility measures. We first calculated inter- and intra-rater reliability measures using the lesion volumes. Based on lesion volumes, the inter-rater reliability was 0.76±0.14, while the intra-rater reliability was 0.84±0.09.

In addition, we also calculated inter-and intra-rater reliability using the dice coefficient (DC), which is a common measurement for image similarity[25]. DC allows us to examine not only if the volumes are similar, but also if the same voxels are being selected as part of the lesion mask or not. This is particularly useful for comparing neuroimaging volumes, such as lesion masks. DC is calculated by the formula:

$$DC = \frac{2|X \cap Y|}{|X| + |Y|}$$

where X and Y represent the voxels from each lesion segmentation, and DC ranges from 0 to 1 (where 0 means there were no overlapping voxels and 1 means that the segmentations were completely the same). An inter-rater DC score was calculated for each manual segmentation by comparing each individual tracer's lesion mask to the rest of the tracers' lesion masks. All inter-rater DC scores were then averaged to obtain one final score for the initial segmentations (average inter-rater DC for first segmentation: 0.745±0.19) and for the secondary segmentations (average inter-rater DC for second segmentation: 0.766±0.16). Furthermore, an intra-rater DC score was calculated for each brain traced by comparing the initial segmentation to the

secondary segmentation for each tracer; these scores were then averaged to obtain a final intra-rater DC score (0.831±0.13).

Trained tracers segmented all lesion masks. In addition, each lesion mask was checked by a separate tracer and changes were made to the lesion mask as needed. Lastly, after the completion of the dataset, lesion masks were checked a second time to ensure correct segmentation and data descriptors. It is important to note that while tracers did go through a training process and segmentations were checked multiple times, this is still a subjective process. Comments regarding the lesion masks can be submitted as issues on the ATLAS Github site (https://github.com/npnl/ATLAS/issues), and we plan to publish updated and expanded versions of this dataset based on feedback and comments from users (see *Usage Notes*).

Usage Notes

The full archived dataset, including both native and standardized datasets, can be found at ICPSR: http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36684. For more information on the data archive, visit the ICPSR website (https://www.icpsr.umich.edu/icpsrweb).

In addition, a standardized, defaced version can be found at FCP-INDI: http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html.

Data is accessible under a standard Data Use Agreement, under which users must agree to only use the data for purposes as described in the agreement. Users of the ATLAS dataset should acknowledge the contributions of the original authors and research labs by properly citing this article and the data repository link from which they accessed the data.

As described above, lesions were segmented using the NITRC open source software MRIcron which can be downloaded from the NITRC website (https://www.nitrc.org/projects/mricron). Users can also quickly and easily view the brains on BrainBox (http://brainbox.pasteur.fr/), an open-source Web application to collaboratively annotate and segment neuroimaging data available online[26]. For additional quick quantification, our group has also created a small toolbox called SRQL (Semi-automated Robust Quantification of Lesions), which includes three features: it uses a semi-automated white matter intensity correction, outputs a report of descriptive statistics on lesions (hemisphere and volume of lesion), and gives users the option to perform analyses in native or standard space (https://github.com/npnl/SRQL)[27]. Finally, any issues or feedback can be submitted on the ATLAS GitHub page (https://github.com/npnl/ATLAS/issues), on which any updates, software, and additional releases will also be announced.

Acknowledgments

We thank Dr. Tony Maguire and Dr. Mauricio Reyes for insightful conversations and would like to acknowledge the following people for their assistance on this effort: Anthony Benitez, Xiaoyu Chen, Cristi Magracia, Ryan Mori, Dhanashree Potdar, Sandyha Prathap. The archiving of this dataset was specifically supported by the NIH-funded Center for Large Data Research and Data Sharing in Rehabilitation (CLDR;

https://www.utmb.edu/cldr) under a Category 2 Pilot Grant (P2CHD06570) and this work was also funded by an NIH K01 award (1K01HD091283).

Author contributions

J.M.A. segmented and reviewed lesions, oversaw the organization to the segmentation process and contributed to the writing and editing of the manuscript. N.W.B. organized, segmented and reviewed lesions. M.S. provided the neuroradiology expertise and information. K.L.I. and H.K. performed data analysis. H.K. also performed data processing and generated the standardized dataset and probabilistic lesion maps. T.A. provided data visualization expertise and generated the figures/videos. J.C., D.S., A.S. J.I., C.J., W.N., D.V. and S.L. segmented and/or reviewed lesions. P.H., B.K., N.K., L.A.-Z., S.C.C., J.L., S.S., L.T.W., J.W., C.W., C.Y. collected and provided the MRI data. M.L., A.P., and A.S. handled the archiving of the data. S.-L.L. conceptualized the study, reviewed lesions, analyzed data, and contributed to the writing and editing of the manuscript.

Competing interests

The authors have no conflict of interest.

Figures

Figure 1. A probabilistic lesion overlap map for the primary lesions from the ATLAS R1.1 dataset. A 3D visualization of the lesion overlap map can be found at https://www.youtube.com/watch?v=Ag5CUsRNY9Q.

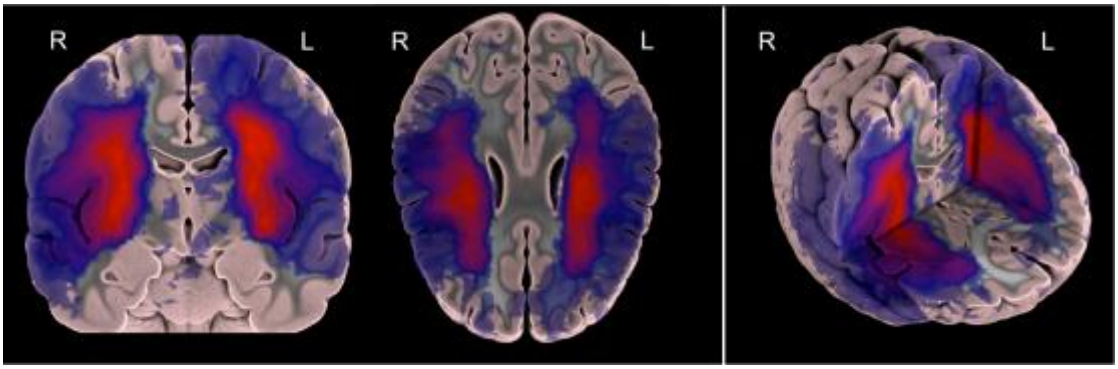


Figure Legends

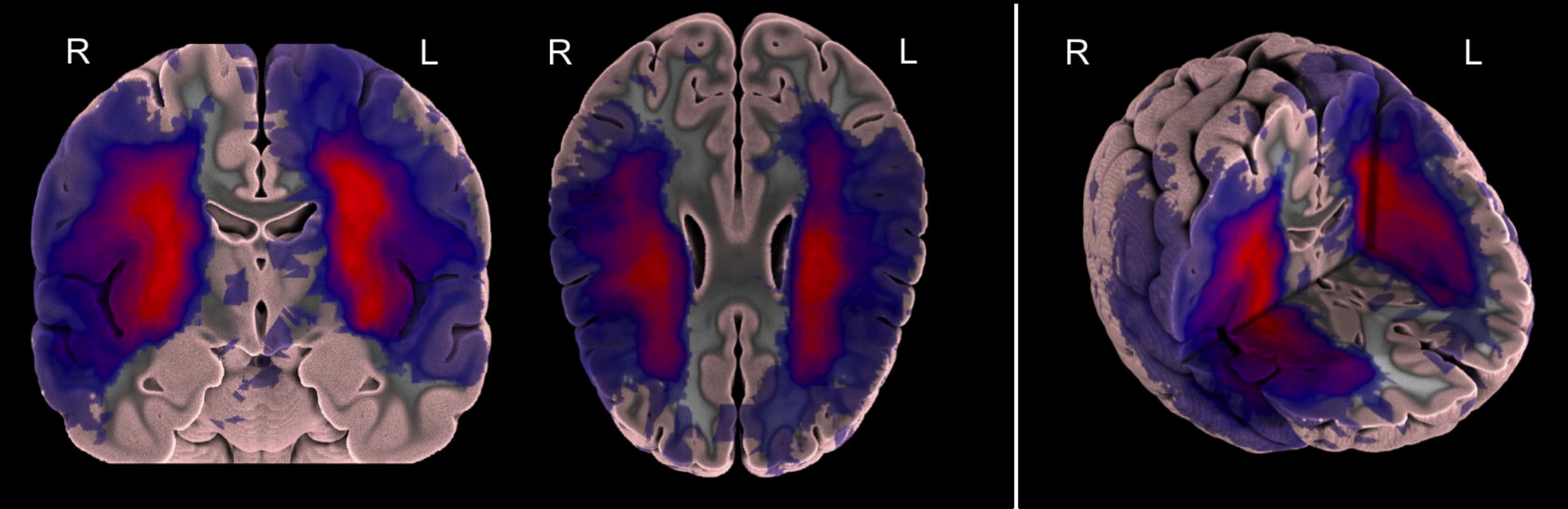
Figure 1. A probabilistic lesion overlap map for the primary lesions from the ATLAS R1.1 dataset.

References

1. Feigin, V. L. *et al.* Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. *Lancet (London, England)* **383,** 245–254 (2014).
2. Kwakkel, G., Kollen, B. J., Van der Grond, J. V. & Prevo, A. J. H. Probability of regaining dexterity in the flaccid upper limb: Impact of severity of paresis and time since onset in acute stroke. *Stroke* **34,** 2181–2186 (2003).
3. Nijland, R. H. M., van Wegen, E. E. H., Harmeling-van der Wel, B. C. & Kwakkel, G. Presence of finger extension and shoulder abduction within 72 Hours after stroke predicts functional recovery. *Stroke* **41,** 745–750 (2010).
4. Marie-Héléne, M. & Cramer, S. C. Biomarkers of recovery after stroke. *Curr. Opin. Neurol.* **21,** 654–659 (2008).
5. Riley, J. D. *et al.* Anatomy of stroke injury predicts gains from therapy. *Stroke* **42,** 421–426 (2011).
6. Prabhakaran, S. *et al.* Inter-individual variability in the capacity for motor recovery after ischemic stroke. *Neurorehabil. Neural Repair* **22,** 64–71 (2007).
7. Stinear, C. Prediction of recovery of motor function after stroke. *Lancet Neurol.* **9,** 1228–1232 (2010).
8. Cramer, S. C. *et al.* Predicting functional gains in a stroke trial. *Stroke* **38,** 2108–2114 (2007).
9. Jongbloed, L. Y. N. Prediction of function after stroke: a critical review. *Stroke* **17,** 765–776 (1986).
10. Zhu, L. L., Lindenberg, R., Alexander, M. P. & Schlaug, G. Lesion load of the corticospinal tract predicts motor impairment in chronic stroke. *Stroke* **41,** 910–915 (2010).
11. Nouri, S. & Cramer, S. C. Anatomy and physiology predict response to motor cortex stimulation after stroke. *Neurology* **77,** 1076–83 (2011).
12. Fiez, J. A., Damasio, H. & Grabowski, T. J. Lesion segmentation and manual warping to a reference brain: Intra- and interobserver reliability. *Hum. Brain Mapp.* **9,** 192–211 (2000).
13. Montaner, J. *et al.* Plasmatic level of neuroinflammatory markers predict the extent of diffusion-weighted image lesions in hyperacute stroke. *J. Cereb. Blood Flow Metab.* **23,** 1403–1407 (2003).
14. Sakamoto, Y. *et al.* Early ischaemic diffusion lesion reduction in patients treated with intravenous tissue plasminogen activator: infrequent, but significantly associated with recanalization. *Int. J. Stroke* **8,** 321–326 (2013).
15. Thomas, R. G. R. *et al.* Apparent diffusion coefficient thresholds and diffusion lesion volume in acute stroke. *J. Stroke Cerebrovasc. Dis.* **22,** 906–909 (2013).
16. Wittsack, H.-J. *et al.* MR Imaging in Acute Stroke: Diffusion-weighted and Perfusion Imaging Parameters for Predicting Infarct Size. *Radiology*

222, 397–403 (2002).

17. Pustina, D. *et al.* Automated segmentation of chronic stroke lesions using LINDA: Lesion identification with neighborhood data analysis. *Hum. Brain Mapp.* **37,** 1405–1421 (2016).
18. de Haan, B., Clas, P., Juenger, H., Wilke, M. & Karnath, H.-O. Fast semi-automated lesion demarcation in stroke. *NeuroImage. Clin.* **9,** 69–74 (2015).
19. Maier, O. *et al.* ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med. Image Anal.* **35,** 250–269 (2017).
20. Rorden, C., Karnath, H.-O. & Bonilha, L. Improving Lesion-Symptom Mapping. *J. Cogn. Neurosci.* **19,** 1081–1088 (2007).
21. Fazekas, F., Chawluk, J., Alavi, A., Hurtig, H. & Zimmerman, R. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *Am. J. Roentgenol.* **149,** 351–356 (1987).
22. Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* **17,** 87–97 (1998).
23. Collins, Louis, D., Neelin, P., Peters, Terrence, M. & Evans, Alan, C. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J. Comput. Assist. Tomogr.* **18,** 192–205 (1994).
24. Cox, R. W. *et al.* A (sort of) new image data format standard: Nifti-1. *Neuroimage* **22,** e1440 (2004).
25. Zou, K. H. *et al.* Statistical validation of image segmentation quality based on a spatial overlap index1: Scientific reports. *Acad. Radiol.* **11,** 178–189 (2004).
26. Heuer, K., Ghosh, S., Robinson Sterling, A. & Toro, R. Open Neuroimaging Laboratory. *Res. Ideas Outcomes* **2,** e9113 (2016).
27. Ito, K., Anglin, J., Kim, H. & Liew, S.-L. Semi-automated Robust Quantification of Lesions (SRQL) Toolbox. *Res. Ideas Outcomes* **3,** e12259 (2017).


R
L
R
L
R
L