Supplemental Methods

Sample selection and behavioural testing:
A total of 56 first-time stroke patients were selected from our PLORAS database\(^1\) that contains hundreds of stroke patients with variable demographics and symptoms, using a retrospective chart review methodology. Selected patients have been scanned twice (n=43) or more (n=13) with time between repeated visits varying between 2 months and 6 years (Supplementary Figure I). They all have visible lesions on MR images, as assessed by a neurologist (APL). Irrespective of the site of the lesion, there was no evidence of other neurological conditions (e.g. dementia, multiple sclerosis), and there were no reports of a second stroke. Across all patients, there was no correlation between age, time post-stroke and time between visits (i.e. Spearman’s rank correlation not significant, p>0.1).

All patients were tested with the Comprehensive Aphasia Test (CAT)\(^2\) that includes subtests for comprehension, repetition, spoken language production, reading aloud and writing. The CAT was designed to tap into many variables known to affect language abilities and it is a suitable test for monitoring changes over time (see discussion in \(^3\)). Out of a total of 128 CAT tests, 81% were administered the same day as the MRI scanning session, 5% within one week, and 14% at a median of 8 weeks from the scanning date.

Data acquisition
A 1.5 Tesla Sonata scanner (Siemens Medical Systems, Erlangen, Germany) was used to acquire all data. Anatomical high-resolution T1-weighted (T1w) images were acquired using a research protocol with an optimized three-dimensional modified driven equilibrium Fourier transform sequence\(^4\) [176 sagittal partitions; image matrix = 256x224; isotropic resolution of 1mm\(^3\); repetition time=12.24ms; echo time=3.56ms; inversion time=530ms].

Data analysis
All analyses were carried out with scripts written in Matlab (The MathWorks, Natick, MA, USA) that called on different processing routines from the statistical parametric mapping (SPM12) software package (Wellcome Trust Centre for Neuroimaging, London, UK). In short, analyses aimed to quantify in an automated way the changes in brain damage and atrophy between repeated scans collected at variable times post-stroke (see illustration of raw T1w data in Supplementary Figure II-A). The estimated change in lesion was quantified in [cm\(^3\)] as well as relative [%] to the total intracranial volume and lesion volume at the earlier time post-stroke. Subsequent statistical analyses looked at associations between the amount of change in brain damage and time between first visit and last visit. We ensured an optimal estimation of the amount of change in a multi-step procedure, as detailed below.

We were confident that the expected volumetric changes can accurately be detected by our optimized acquisition protocol and segmentation methods. This is because the very few studies that previously investigated long-term volumetric changes have observed large...
detectable changes. For instance, using manual tracing on CT scans of 12 patients, Naeser and colleagues\(^5\) showed a significant expansion in lesion borders after 5 years post-stroke, with an average increase in lesion size of 3.3%. In 10 stroke patients, Kraemer and colleagues\(^6\) reported an average brain volume shrinkage of >110cm\(^3\) after 1-4 years post-stroke due to marked shrinkage of the cerebral gyri adjacent to the infarction, atrophy in remote regions, and enlargement of the lateral ventricles.

**Voxel-by-voxel matching between repeated scans.** The first step was to match voxel-by-voxel the two longitudinal anatomical images in terms of spatial location and signal intensity.\(^7\), \(^8\) This is necessary because, at each time-point, the patient’s head will be positioned slightly differently in the head coil, acquired at variable signal intensity, and affected by time-varying B0-inhomogeneities (i.e. a smooth and spatially varying bias field). For each scan, a bias field correction was performed using the unified normalization-segmentation algorithm\(^9\) that incorporates a model for smooth intensity variations. This resulted in bias-free T1w scans that were segmented into different tissue classes in the native space (gray matter, white matter, cerebrospinal fluid, skull and scalp) using the New Segment tool of SPM12.

To estimate brain size, a mask of the whole brain was defined as the sum of gray matter, white matter and cerebrospinal fluid tissue classes, thresholded at 0.9. To fill in the holes after thresholding, a large disk was used as a structuring element in a morphological “closing” operation.\(^10\) The whole brain binary mask of each patient was applied to the bias-free T1w scan to generate a skull-stripped brain scan and its total size was used as an approximation of the total intracranial volume,\(^11\) which is needed to adjust for individual differences in cranial/brain size in statistical analyses.\(^12\), \(^13\) Across our 56 patients, total intracranial volume was 1493cm\(^3\) on average (SD=148cm\(^3\)) at the first visit and 1477cm\(^3\) on average (SD=128cm\(^3\)) at the last visit, which is comparable to previous literature.\(^14\) These estimates at first and last visit were almost perfectly correlated (Pearson correlation r=0.98, see Supplementary Figure II-B), showing the robustness of the segmentation procedure when dealing with T1w scans with large lesions.

Signal intensities between each pair of the bias-free and skull-stripped T1w scans were matched using a standardization procedure\(^15\) based on the estimated signal intensities for each tissue class from the unified normalization-segmentation algorithm (c.f. equations [23,25] in\(^9\)) that modelled T1w signal intensities with a mixture of Gaussians. Last but not least, for each patient, the resulting T1w scans were spatially co-registered to the earliest scan, at the first time point, using a rigid body registration by maximization of mutual information that ensures sub-voxel accuracy.\(^16\), \(^17\) A normalized cross-correlation was used as an objective function in the co-registration procedure that was shown to be optimal for intra-modal registration.

**Standardized difference estimation between repeated scans.** The second step was to quantify the longitudinal changes in brain damage at the individual patient level. Using the T1w scans that had been matched in location and intensity, we subtracted the scan at the first time point from the scan at the later time point. This resulted in a whole brain difference image that coded the relative difference in intensity between repeated scans on a voxel by voxel basis. This difference image can be positive (intensity lower later than earlier) or negative, with the expectation here of mainly identifying positive differences due to lesion growth\(^5\) and atrophy;\(^6\), \(^18\), \(^19\) see illustration in Supplementary Figure III.

To control for differences in signal intensity that might be caused by random MR noise only,\(^20\) we standardised the difference score at each voxel, assuming a normal distribution of MRI noise across space and time. The standardized difference image was created by dividing
the difference in signal intensity at each voxel by the summed standard deviation of
differences that was computed as the square root of the sum of within-scan variances of the
white matter tissue as estimated by the mixture of Gaussians model in the unified
normalization-segmentation algorithm above. The standardized difference image was then
thresholded at 1.96 (eq. two-tailed p<0.05), yielding a whole-brain binary image of
“longitudinal change” that showed significant difference between two repeated T1w scans.
Although all T1w scans were acquired here with the same protocol (i.e. same 1.5T scanner
and the same high-resolution sequence), we cannot rule out that some residual
differences may exist between repeated T1w scans (e.g. changes in contrast). For instance, such subtle
residual differences might result from differences in the way the MR radio frequency
excitation interacted with variable head positions in the head coil. However, such differences
are unlikely to be spatially specific and hence using variable lesion sites over many patients
would make their contributions negligible.
To differentiate between longitudinal changes within- and outside the lesion, the longitudinal
change images were compared to the patient’s lesion image from the first scan. The lesion
boundaries were identified using an automated lesion identification algorithm operating in
MNI space\textsuperscript{21} that was then warped back into the native space. We preferred to use a fully
automated lesion identification procedure rather than manual lesion-tracing to avoid
introducing operator-dependent errors. All lesions identified by our automated procedure at
the first time point were checked by a neurologist (APL) who was blind to the aim of the
current study and had no access to the repeated scans of each patient. Practically, each
patient’s T1w scan was segmented into different tissue types using a modified unified
segmentation-normalization procedure that incorporates an explicit extra prior for the lesion.
The damaged tissue is assigned to the extra class, which means that the probability of being
gray or white matter is reduced at the site of the lesion. The normalized gray and white matter
tissue classes of each patient were then spatially smoothed and subsequently compared
voxelwise to a set of smoothed gray and white matter images of healthy matched controls
using an optimized outlier detection algorithm based on fuzzy clustering.\textsuperscript{21} The outlier
detection algorithm generates a 3D definition of each patient’s lesion in MNI space.
Importantly, this automated procedure ensured minimal bias when assessing longitudinal
changes because it was agnostic to the lesion information in the other repeated T1w scans
(i.e. only applied on the first T1w scan). It is also worth noting that all volumetric changes
were estimated here directly from the standardized difference image that was generated prior
to lesion identification. Using our automated procedure, lesion volume at the first scan ranged
from 0.5cm\textsuperscript{3} to 367cm\textsuperscript{3} across patients (median=43cm\textsuperscript{3}, see Supplementary
Figure II-C).
The degree of change in lesion was defined as the set of voxels in the longitudinal change
image that was outside the initial lesion boundaries, but morphologically connected to it at up
to 10mm. The latter helped to minimise the contribution of atypical ventricular dilatation
(secondary lesion effects) in patients with very large lesions (e.g. >200cm\textsuperscript{3}). The degree of
change outside the lesion was defined as the longitudinal changes in the rest of the brain
(excluding infarcted voxels). Such effects are likely to be due to atrophy such as sulci
widening and dilated ventricles, see illustration in Supplementary Figure IV. The expectation
here was that the shrinkage in post-stroke brains would be greater than that previously
reported for typical age-related atrophy.\textsuperscript{22, 23} In patients with unilateral damage (53 out of 56
patients), the rate of atrophy was also computed separately for ipsi-lesional and contra-
lesional hemispheres (excluding voxels located at the inter-hemispheric fissure). Our
estimates are expressed at the global level (i.e. as a whole-brain atrophy rate in [% /year] after
excluding infarcted voxels) and thus we did not look specifically at other focal but subtle
structural changes over the time course of recovery that may include regional changes in
atrophy,\textsuperscript{24} gray matter volume\textsuperscript{25} and cortical thickness.\textsuperscript{26}
**Statistical analysis.** Here we used nonparametric tests as our variables were skewed. The focus of our analysis was to assess the degree of association between changes in brain damage and the time between visits using the cross-sectional data over our 56 patients. A Spearman partial rank correlation analysis was used to test for significant associations while controlling for many factors at the first visit including age, lesion volume, time post-stroke and total intracranial volume. We repeated the same partial correlation analysis after normalizing longitudinal changes in brain damage [in %] to the total intracranial volume while controlling for age, lesion volume and time post-stroke (see discussion in 12, 13). We derived accurate estimates of the median rate of lesion growth with time between repeated visits [cm$^3$ per year] using the longitudinal data available in the 13 patients who were scanned three or four times. Cross-sectional plots of changes in language test scores from the CAT over time were also generated. Last but not least, using Spearman's rank correlation coefficient, we also assessed the impact of such volumetric changes on long-term recovery of language skills as derived from the CAT scores.

**Supplemental Results**

1- Overall, the amount of brain shrinkage (i.e. lesion growth and atrophy) was significantly different from zero (Wilcoxon signed rank test: p<0.001) but varied considerably across patients, as illustrated in Figure 1 of the main manuscript. The significant relationship between lesion change and time between repeated visits also survived adjustment for the total intracranial volume (Spearman partial rank correlation rho=0.53, p<0.001).

2- Although our sample was not counterbalanced with respect to gender (25% females versus 75% males), differences between males and females were not expected to influence our volumetric estimates. Over all patients, female subjects tended to have smaller total intracranial volume than male subjects (as illustrated in Figure II-B, Mann–Whitney U test: p=0.005) and they were slightly (p=0.01) younger (51.3 years, SD=12.5) than males (61.5 years, SD=9.5); however, differences between males and females were critically not significant on any of the variables that were associated with longitudinal volumetric changes, including the gap between first and last visit (Mann–Whitney U test: p=0.9), lesion volume (p=0.61), and time post-stroke (p=0.30). Most importantly, our data showed that the rate in lesion growth, in [%] after adjusting for differences in total intracranial volume, was not significantly different between females and males (p=0.39).

3- As expected (e.g. see 5,6), despite significant brain shrinkage, long term recovery continued as demonstrated here by the improved CAT scores at later visits in the majority of patients (see Supplementary Figure V for more details). The improvement was observed for several tasks including, for instance, object naming (Wilcoxon signed-rank test p=0.01). Moreover, there was no significant relationship between changes in the behavioural CAT scores and the rate of lesion growth (at p<0.05 Bonferroni-corrected for multiple statistical testing); see the exact values of the Spearman's rank correlation coefficients in Supplementary Figure V.
Supplemental References


Figure I: histograms showing the number of patients in our sample according to their age (top, in years), time post-stroke (middle, in months) and time between repeated visits (bottom, in months).
**Figure II:** (A) illustrations of visible changes to a unilateral left lesion in a patient with 4 repeated T1w scans. (B) a scatter plot of the total intracranial volume at earlier visit versus last visit of all patients (square marks = female patients, circle marks = male patients). (C) histogram showing the number of patients according to lesion size at first visit.
**Figure III:** Illustration of the difference image between repeated T1w scans of two patients. The original T1w scans of each patient were first matched in terms of signal intensity and location (signal standardization and spatial registration) and then subtracted to generate a difference image. Bright regions in the difference image code voxels where T1w intensity at later time was lower than T1w intensity at earlier time.
**Figure IV:** Illustration of atrophy, here in terms of ventricular dilatation and sulci widening in four different patients. Bright regions in the difference image code voxels where T1w intensity at later time was lower than T1w intensity at earlier time.
Figure V: each subplot shows the raw CAT scores at earlier visit against the raw CAT scores at later visit. Each patient is represented by a black circle with a radius proportional to the amount of change in lesion size. Dots above the diagonal line represent higher CAT scores at later than earlier visit. Max possible scores (performance at ceiling) are illustrated by a dotted line. The Rho values represent the Spearman's rank correlation coefficient between the changes in the CAT scores for each subtest and the rate of lesion growth: rho=0.08 (p=0.53) for spoken word comprehension, rho=-0.15 (p=0.26) with word reading, rho = 0.21 (p=0.11) for word repetition, rho = 0.17 (p=0.21) with fluency, rho=0.30 (p=0.03) with object naming, and rho=0.31 (p=0.02) with picture description.