Bayesian lesion-deficit inference with Bayes factor mapping: key advantages, limitations, and a toolbox

Christoph Sperber , Laura Gallucci , Stefan Smaczny , Roza Umarova

PII: S1053-8119(23)00154-4
DOI: https://doi.org/10.1016/j.neuroimage.2023.120008
Reference: YNIMG 120008

To appear in: *NeuroImage*

Received date: 10 November 2022
Revised date: 6 March 2023
Accepted date: 7 March 2023

Please cite this article as: Christoph Sperber , Laura Gallucci , Stefan Smaczny , Roza Umarova , Bayesian lesion-deficit inference with Bayes factor mapping: key advantages, limitations, and a toolbox, *NeuroImage* (2023), doi: https://doi.org/10.1016/j.neuroimage.2023.120008

Highlights

- Lesion-deficit inference is so far dominated by frequentist statistical mapping
- We evaluated Bayesian lesion-deficit inference in silico and in real deficits
- Bayesian inference was statistically more liberal than frequentist mapping
- It transparently handles situations with low statistical power
- It complements the lesion mapping method portfolio with unique advantages

Bayesian lesion-deficit inference with Bayes factor mapping: key advantages, limitations, and a toolbox

Christoph Sperber[1], Laura Gallucci[1], Stefan Smaczny[2], Roza Umarova[1]

[1] Department of Neurology, Inselspital, University Hospital Bern, University of Bern, Bern, Switzerland

[2] Centre of Neurology, Hertie-Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

Corresponding Author:
Christoph Sperber
Universitätsklinik für Neurologie, Inselspital
Freiburgstr. 16
3010 Bern
Switzerland
e-mail: christoph.sperber@extern.insel.ch

Running Title: Bayesian lesion-deficit inference

Abstract

Statistical lesion-symptom mapping is largely dominated by frequentist approaches with null hypothesis significance testing. They are popular for mapping functional brain anatomy but are accompanied by some challenges and limitations. The typical analysis design and the structure of clinical lesion data are linked to the multiple comparison problem, an association problem, limitations to statistical power, and a lack of insights into evidence for the null hypothesis. Bayesian lesion deficit inference (BLDI) could be an improvement as it collects evidence for the null hypothesis, i.e. the absence of effects, and does not accumulate α-errors with repeated testing. We implemented BLDI by Bayes factor mapping with Bayesian t-tests and general linear models and evaluated its performance in comparison to frequentist lesion-symptom mapping with a permutation-based family-wise error correction. We mapped the voxel-wise neural correlates of simulated deficits in an in-silico-study with 300 stroke patients, and the voxel-wise and disconnection-wise neural correlates of phonemic verbal fluency and constructive ability in 137 stroke patients. Both the performance of frequentist and Bayesian lesion-deficit inference varied largely across analyses. In general, BLDI could find areas with evidence for the null hypothesis and was statistically more liberal in providing evidence for the alternative hypothesis, i.e. the identification of lesion-deficit associations. BLDI performed better in situations in which the frequentist method is typically strongly limited, for example with on average small lesions and in situations with low power, where BLDI also provided unprecedented transparency in terms of the informative value of the data. On the other hand, BLDI suffered more from the association problem, which led to a pronounced overshoot of lesion-deficit associations in analyses with high statistical power. We further implemented a new approach to lesion size control, *adaptive lesion size control*, that, in many situations, was able to counter the limitations imposed by the association problem, and increased true evidence both for the null and the alternative hypothesis. In summary, our results suggest that BLDI is a valuable addition to the method portfolio of lesion-deficit inference with some specific and exclusive advantages: it deals better with smaller lesions and low statistical power (i.e. small samples and effect sizes) and identifies regions with absent lesion-deficit associations. However, it is not superior to established frequentist approaches in all respects and therefore not to be seen as a general replacement. To make Bayesian lesion-deficit

inference widely accessible, we published an R toolkit for the analysis of voxel-wise and disconnection-wise data.

Keywords

Lesion-symptom mapping; VLSM; verbal fluency; voxel; disconnection; stroke

Abbreviations:BF, Bayes factor; BLDI, Bayesian lesion-deficit inference; NHST, Null hypothesis significance testing; VLSM, Voxel-based lesion-symptom mapping

1 Introduction

Lesion-deficit inference identifies brain regions in which lesions are associated with the occurrence or persistence of cognitive and behavioural deficits. On the one hand, this information can be used in diagnosis or prognosis. On the other hand, it provides crucial information for mapping the functional architecture of the human brain because, unlike correlational methods such as fMRI or EEG, it can identify brain regions that are critical for cognitive function (Rorden and Karnath, 2004). Since the introduction of univariate statistical mapping (Bates et al., 2003; Rorden and Karnath, 2004), this approach has dominated the field of lesion-deficit inference and has been used by hundreds of studies (Karnath and Rennig, 2017). The basic principle is to statistically evaluate the association between the lesion status of each imaging feature, such as each voxel in a normalised imaging space, and a measure of a post-stroke deficit. The method is rooted in the framework of statistical parametric mapping (Friston et al., 1994) and is therefore flexible and elegantly simple. Voxel-wise lesion-deficit inference has been used with a plethora of statistical tests, such as t-tests (Bates et al., 2003), χ^2-tests (Karnath et al., 2004), and non-parametric tests (Rorden et al., 2007) and can easily be transferred to various imaging data of pathological brain aberrations, including region-wise lesion load (Behroozmand et al., 2022), functional network disconnectivity (Boes et al., 2015), white matter disconnection (Umarova et al., 2014), and indirectly estimated structural disconnection (Sperber et al., 2022).

However, the success of univariate statistical lesion-deficit inference is overshadowed by several key problems. First is the *multiple comparison problem*. The repeated use of statistical tests (e.g. in each voxel of the brain) massively inflates α errors. This α error inflation can satisfyingly be countered by multiple comparison correction, for example by permutation methods (Nichols and Holmes, 2002; Mirman et al., 2018). Second is the *association problem.* The vascular supply of the brain is

structured hierarchically which defines typical lesion patterns (Zhao et al., 2020). Imagine a deficit that arises after damage to only one specific brain region. Lesions that damage this brain region often damage other brain regions that are supplied by a vascular branch within the same supply hierarchy (see Figure 1A for an illustration). Therefore, the deficit is also associated with damage to these other regions. The results of lesion-deficit mapping were found to be biased by these typical associations in lesion anatomy (Mah et al., 2014; Pustina et al., 2018; Sperber et al., 2019), causing an overshoot of lesion-deficit associations beyond the neural correlates of a deficit. This appears to be a general and currently unsolved limitation of any lesion-deficit approach, be it univariate or multivariate (Sperber et al., 2019; Sperber, 2020). Third, various properties of lesion-deficit data cause a *statistical power problem*. Statistical power is the probability of correctly rejecting the null hypothesis (i.e. identifying true lesion-deficit associations) and it depends, in addition to effect size, on sample size and group sizes. The sample size in neuropsychological studies with stroke patients is often small and the data structure is not optimal. Consider the case of binary voxel-wise lesion data, where a voxel is either lesioned or intact. Here, a statistic usually compares whether the severity of the deficit differs between patients with damage in the voxel and patients without damage in the voxel. Statistical power is highest when both groups are equally sized (Kimberg et al., 2007). However, group sizes are often highly unequal in stroke patient samples and vary across brain regions due to the stroke anatomy (see Figure 1B). This problem is further amplified in samples with small lesions (see Figure 1C). A small average lesion size could be particularly problematic in contemporary studies as, thanks to innovations in intravenous thrombolysis and mechanical thrombectomy, brain tissue can be saved and more patients survive a stroke with only minor lesions and deficits (Bhaskar et al., 2018). Another limitation of statistical power is introduced by the partial injury of neural correlates within networks (Figure 1D; Rorden et al., 2009). This issue arises whenever lesions damage only parts of a neural correlate, which is especially problematic with more complex neural correlates, such as large-scale cognitive networks. For example, when a deficit arises after damage to either of two regions, a patient with the deficit after damage to one region serves as a statistical counter-example that the deficit also arises after damage to the other region. The statistical power to also detect the second region as the deficit's neural correlate is then reduced.

Figure 1 Limitations in LSM near here, 2 columns

Until now, lesion-deficit inference was strongly dominated by frequentist null hypothesis significance testing (NHST). With NHST, we can only learn something about data if the null hypothesis h_0 is rejected. Any result above the α level is not informative as it only indicates the absence of evidence for an effect, but not evidence for the absence of an effect (Dienes, 2014; Keysers et al., 2020). Bayesian inference follows another statistical concept that can provide evidence for either h_0 or the alternative hypothesis h_1. The basic principle in Bayesian inference is that we start with a *prior probability* for the presence of an effect. A prior probability can be based, for example, on subjective beliefs or previous studies, or it can be an uninformative default prior. Next, we look at the evidence provided by our research data and update the prior probability accordingly, thereby computing the *posterior probability*. Even though Bayesian methods are currently trending in neuroscience (Gallistel, 2009; Wagenmakers et al., 2018; Keysers et al., 2020), they are still a rare sight in lesion-deficit inference. A few implementations either used noninformative default priors to map the posterior probability (Chen and Herskovits, 2010), Bayesian statistics on low-dimensional lesion imaging data (Bonkhoff et al., 2021; Bonkhoff et al., 2022) or included Bayesian network analyses (Chen et al., 2008; Duering et al., 2014; Arnoux et al., 2018) that go beyond the simple Bayesian inference framework. A few more studies mapped the Bayes factor (Achilles et al., 2017; Ulrichsen et al., 2021). The Bayes factor quantifies the strength of the evidence provided by data for the evaluation of h_0 versus h_1. Bayes factor mapping has the advantage that it is, to a large degree, independent of the prior probability and, instead, informs us only about the evidence provided by the data. Inference with the Bayes factor also has a practical advantage: for the most popular frequentist statistical tests, including t-tests, ANOVAs, rank tests, and correlation analysis, counterparts utilising the Bayes factor exist (Rouder et al., 2012; Wetzels et al., 2012; Wetzels and Wagenmakers, 2012; Keysers et al., 2020; van Doorn et al. 2020). Hence, inference using the Bayes factor should be easily accessible even to scientists that are only familiar with the common tests for NHST. And, of importance to lesion-deficit inference, the Bayes factor can be mapped within the existing framework of voxel-wise statistical mapping.

Bayesian methods might better handle the difficulties in lesion-deficit inference. First, it solves the *multiple comparison problem* because it simply does not

generate α-errors. Second, Bayesian inference could be a transparent and informative approach to handling the *statistical power problem* in lesion-deficit inference. A switch from NHST to Bayesian inference would not necessarily increase our ability to gather evidence for weak and subtle effects. However, Bayesian inference could transparently differentiate brain areas for which our data prove the null hypothesis and brain areas for which our data do not provide any evidence. On the other hand, the statistical power problem in lesion-deficit inference might go beyond aspects that Bayesian inference can transparently handle. If statistical power is low due to small group- or sample sizes – a central issue in lesion-deficit data – effects can be overestimated (Button et al., 2013; Gelman and Carlin, 2014). Thus, if we too much rely on the advantage of Bayesian lesion-deficit inference to evaluate the presence of lesion-deficit associations even under the typically difficult conditions of lesion data, we may still overlook biases. Further, the *association problem* remains. How well can Bayesian inference indicate the absence of lesion-deficit associations if these are inflated within the proximity of the neural correlates of a deficit?

In the present study, we evaluated how well Bayes factor mapping can handle the challenges in lesion-deficit inference. We compared classical frequentist lesion-symptom mapping with null hypothesis significance testing to a corresponding Bayes factor mapping approach. First, we compared the methods for voxel-wise mapping with in silico data for which the ground truth of a neural correlate was precisely known and specifically chosen to evaluate the challenges of lesion-deficit data. Second, we mapped and compared voxel-wise data and structural disconnection data for both methods in real-world data on phonemic verbal fluency and constructive ability.

2 Methods

2.1. Samples and lesion data

In the present study, we used two data samples. The first sample included 169 stroke patients recruited for the CogStroke study at the University Hospital Bern (ClinicalTrials.gov Identifier: NCT05653141). All patients suffered a first-ever ischemic stroke in the anterior circulation. Detailed recruitment criteria and demographic and clinical data are reported in the supplementary. The second sample included 131 stroke patients from a publicly available sample within the LESYMAP software (Pustina et al., 2018). For the first study part, the in silico validation of

Bayes Factor mapping, we included both datasets to create a large sample of 300 stroke patients. In the second study part, the real-world comparison of Bayes Factor mapping and statistical mapping with significance testing, we investigated a sub-sample of 137 patients from the CogStroke study, which were non-aphasic and completed the neuropsychological assessment. All patients of the CogStroke study gave consent for scientific data use in accordance with the revised Declaration of Helsinki and the study was approved by the local ethic committee.

Lesion maps, i.e., binary topographies of the structural lesion visible on clinical imaging, were generated as described previously (Umarova et al., 2011) or were already available for the public data set. The creation of each lesion map included i) the delineation of the lesioned area on MRI or CT, and ii) the normalisation of the clinical image and the delineated map to MNI space using procedures appropriate for diseased brains. We resliced all lesion maps to a 1x1x1 mm^3 space. In the in silico validation, we mirrored lesions in the right hemisphere along the sagittal midplane to the left hemisphere to create a unilateral lesion sample with a high lesion overlap and, thereby, potentially high statistical power with the maximum sample size. Overlap topographies of all 300 lesions are shown in Figure 2, and of the subsample with 137 lesions in Figure 3. To illustrate the flexibility of Bayesian inference, we performed an additional connectome-based analysis in the sample with 137 lesions and mapped the structural disconnection underlying deficits in constructive ability and phonemic verbal fluency. Detailed methods and results are reported in the supplementary.

Figure 2 Lesion Overlaps Sample 1 near here, 2 columns

Figure 3 Lesion Overlaps Sample 2 near here, 2 columns

2.2. Lesion-deficit inference with frequentist tests and Bayes factor mapping

As a frequentist reference method with null hypothesis significance testing (NHST), we chose voxel-wise statistical lesion-symptom mapping in NiiStat (https://www.nitrc.org/projects/niistat). NiiStat uses voxel-wise general linear models to investigate lesion-deficit associations. An advantage of general linear models is their flexibility. They can be used, for example, on lesion data with binary voxel information (lesioned/intact), where they are equivalent to lesion-symptom mapping

with t-tests (as in Bates et al., 2003), but they can as well be used on continuous voxel- or region-wise data. We controlled for multiple comparisons with a maximum statistic threshold at an α-level of 0.05, which provides an approximately exact family-wise error correction (Nichols and Holmes, 2002) and is often considered to be a gold standard for multiple comparison correction in brain mapping (Karnath et al., 2018; Pustina et al., 2018). We set the number of permutations to 1000 in the in silico study part to save computational resources and to 5000 in the study part on real-world deficits.

We implemented Bayesian lesion-deficit inference (BLDI) by Bayes factor (BF) mapping with the BayesFactor package v0.9.12 (Morey et al., 2018) in R (R Core Team, 2022). We chose two different approaches – BLDI either with a t-test or a general linear model. A t-test is the intuitive first choice to compare a continuous variable (the deficit) between two groups (patients with a lesion in a voxel vs. patients without a lesion in a voxel). However, as for the frequentist null hypothesis significance tests in NiiStat, general linear models could achieve the same results as t-tests, but with more flexibility, allowing the inclusion of covariates and application in continuous imaging data. In each voxel, we computed i) a Bayesian two-sample t-test (Rouder et al., 2009) comparing the severity of a deficit between all patients with a lesion to the voxel vs. all patients without a lesion to the voxel, and ii) a general linear model of the binary lesion status on the deficit score with a mixture of g priors (Liang et al., 2008; Rouder and Morey, 2012) to test the model with the voxel's lesion status against the intercept-only model. We then mapped the resulting voxel-wise BFs back into MNI brain space. We evaluated the results of BF mapping by the conventions set by Wagenmakers and colleagues (2018) and noted moderate evidence for h1 with a BF > 3, strong evidence with a BF > 10, and very strong evidence with a BF > 30. Vice versa, we noted moderate/strong/very strong evidence for h0 at a BF < 1/3 / BF < 1/10 / BF < 1/30. For simplification of visualisation and analysis, we omitted the 'extreme evidence' category.

In the in silico analyses, we only analysed voxels that were damaged in at least five patients. In the analyses of real-world neuropsychological deficits, we lowered this threshold to 4 to include a larger part of the left hemisphere, which was damaged less often and by smaller lesions. Such a threshold value is in principle unavoidable since it prevents invalid statistical tests (like a t-test of 50 vs. 0 or 49 vs. 1 patients) and situations with extremely little data variance. Therefore, it is commonly included

in lesion-deficit inference, while the exact threshold varies to some degree across the literature (see Sperber and Karnath, 2017) and is chosen to some degree arbitrarily. We subjected BLDI and the NHST method to the same threshold and, additionally, report the amount of excluded voxels in the results. The implications of using such a threshold in the context of BLDI and lesion-deficit inference in general are further elaborated on in the discussion.

2.3. Validation of Bayesian lesion-deficit inference – In silico validation

The first study part evaluated BF mapping for lesion-deficit inference in silico, i.e., by computational simulation of deficits under well-controlled and transparent conditions. We adapted the in silico concept from several previous studies that evaluated newly implemented lesion-deficit inference methods (e.g., Zhang et al., 2014; Pustina et al., 2018). The basic principle is to arbitrarily define any brain area, such as one or multiple regions taken from a brain atlas, as the neural correlates of a deficit. In a sample of patients with normalised lesion maps, a deficit is then simulated based on the overlap of each patient's lesion map and the deficit's neural correlate. In other words, the severity of a simulated deficit is 'caused' by a patient's damage to the neural correlate. As we have chosen the neural correlate ourselves, we have perfect knowledge about the organisation of the neural correlate which, in turn, allows us to evaluate the performance of brain mapping methods. Precise and valid lesion-deficit inference should identify the neural correlates of the simulated deficits and any systematic deviations hint at limitations of the method. Importantly, the simulated deficits still originate from real stroke data and are thereby affected by the problems of lesion-deficit inference, such as unequal groups and spatially inflated associations due to stroke anatomy. The detailed concept is illustrated in Figure 4 and was as follows: We used parcels from the Automatic Anatomical Labelling atlas (AAL; Tzourio-Mazoyer et al., 2002) to define the neural correlates of simulated deficits. In a pre-analysis, we identified all regions in the atlas that were damaged in at least 30 out of 300 patients to ensure that simulated deficits contain at least some variance. These regions might have contained parts with a lower lesion coverage than 30 out of 300 patients, which were ignored in our simulation. Forty-five regions fulfilled this criterion (see online materials for details). In each simulation, we picked one of these areas and assessed its overlap with the lesion map of each patient. The patient's deficit was then simulated from this overlap based on a sigmoidal function (see

supplementary for details). This measured overlap was also subjected to random normal noise to make deficits more representative of actual deficits (see Pustina et al., 2018). The standard deviation of the normal noise was equivalent to an overlap of 10% of the region. We varied simulation parameters across three experimental conditions: A) In the first condition, we meant to investigate the impact of the sample size which is one of the most central factors to statistical power and the estimation of effects. We either picked the entire sample of 300 patients or randomly picked subsamples of 100 or 50 patients. B) In the second condition, we wanted to evaluate the impact of small lesion size on lesion-deficit inference. We applied a median split for lesion size in the total sample of 300 patients and randomly picked 100 patients from the 150 patients below the median, i.e. half of the patients with smaller lesions. C) In the third condition, we wanted to find out how lesion-deficit inference performs in the case of partial injury of a neural correlate. Therefore, we included a second brain region in the simulation. Each of the 45 brain regions was paired with another, non-adjacent of the 45 brain regions. The second region was chosen randomly and the region pair was the same across all experimental conditions. The deficit was then simulated based on the brain region that overlapped the most with a patient's lesion map. Hence, damage to only one of the two regions sufficed to cause a deficit, which corresponded to the previously described partial-injury problem (Rorden et al., 2009; Sperber et al., 2019). The second region was then ignored in the analysis, i.e. hits and misses were only registered for the first, but not the second region. Again, we randomly picked 100 patients out of the total sample for this condition. In summary, we performed 3 x 45 simulations in condition A (45 each for three sample sizes with n=300, n=100 and n=50), 45 simulations in condition B, and 45 simulations in condition C. The sub-samples of 50 or 100 patients were the same across the different methods (i.e. BLDI with general linear model versus BLDI with t-test versus VLSM with NHST), but they were randomly re-sampled for each simulation (i.e. between Simulation 1 with region 1, Simulation 2 with region 2, etc.).

The simulated deficits were then mapped either by BF mapping with Bayesian t-tests, Bayes factor mapping with general linear models, or a frequentist voxel-wise statistical mapping with general linear models and maximum statistic permutation correction. We then compared each resulting topography with the brain region that was used in the simulation, i.e. the ground truth for the deficit's neural correlate. We performed statistical analyses in R statistics and report the results of null hypothesis

significance tests at $\alpha = 0.05$ with Bonferroni correction and the Bayesian counterpart in the R BayesFactor package.

Figure 4 study concept near here, 2 columns

2.4. Validation of Bayesian lesion-deficit inference – Real-world applications

The second study part compared statistical lesion mapping with NHST and BLDI in actual post-stroke cognitive deficits. We neuropsychologically examined a sample of 137 patients with first-ever ischemic stroke within 10 days of stroke onset in the patient's native language. We assessed constructive ability as included in the CERAD battery (Morris et al., 1989) and phonemic verbal fluency as included in the Regensburg verbal fluency test (Aschenbrenner et al., 2000). For constructive ability, patients were instructed to copy four different line drawings of increasing complexity (a circle, a diamond, two intersecting rectangles, and a cube). Their performance was rated on an 11-point scale according to 2-4 pre-defined evaluation criteria for each shape. Phonemic verbal fluency was assessed by asking the patient to generate as many words with a specific starting letter as possible within 60 seconds. Out of a larger neuropsychological test battery, we chose these two tests for the different degrees of clinical variance they provided, which should result in different effect sizes. We converted the measures into Z-scores according to healthy norm data corrected for age, sex, and education. Patients displayed severe deficits in constructive ability, with an average Z-value in comparison to norm data of $Z = -1.18$. Deficits in verbal fluency were less pronounced, with an average value of $Z = -0.39$. The results, including the extent and peak of statistical maps, verified our assumption on the different degrees of clinical variance.

We mapped neural correlates of both deficits after Z-value standardisation either with frequentist voxel-wise statistical mapping or with Bayes factor mapping by Bayesian general linear models. The evaluation focussed on the correspondence of results in favour of hypothesis h1 and the extent of areas or disconnections for which BLDI can provide new insights by gathering evidence for h0.

3 Results

3.1. In silico validation

The main results of the in silico study are shown in Figure 5, with additional details reported in supplementary tables 1-3. Example results comparing frequentist voxel-based lesion-symptom mapping (VLSM) and Bayesian lesion-deficit inference (BLDI) with Bayesian general linear models are shown in Figure 6. We referenced all outcome variables, such as false positives, true positives, etc., to the total number of positives/negatives in per cent to make them comparable between simulations with different numbers of positives and negatives. We restricted statistical analyses to possible main effects that were of relevance to our conclusions. Repeated frequentist tests across all conditions were corrected for multiple comparisons by Bonferroni correction at an overall $\alpha = 0.05$ and corrected p-values are reported.

A first non-statistical, numerical investigation of the results of VLSM and BLDI suggested that our experimental conditions worked as intended. With larger sample sizes, more areas of the brain were tested, and more true and false evidence for h1 was created. With only small lesions in condition B, the tested areas as well as evidence for h1 were smaller than with unselected lesions. With a second simulation region in condition C, we found slightly less true evidence for h1.

We first statistically investigated i) the proportion of true positive results in favour of h1, i.e. voxels with correct evidence for h1 among all voxels for which h1 was true and ii) the proportion of false positive results in favour of h1, i.e. voxels with false evidence for h1 among all voxels for which h0 was true. We used paired t-tests to compare frequentist VLSM to BLDI with t-tests, VLSM with BLDI with general linear models, and the two BLDI methods. Across all conditions, BLDI with t-tests found more true positives (all $p < 0.0001$; all BFs > 1170) and more false positives (all $p < 0.0001$; all BFs > $1.1*10^7$) than VLSM. Likewise, BLDI with GLMs found more true positives (all $p < 0.0001$; all BFs > 1100) and more false positives (all $p < 0.0001$; all BFs > $1.0*10^7$) than VLSM. The comparison of the two BLDI methods was inconclusive about simulation condition A with 300 subjects, for which no evidence for a difference in true positives could be found ($t(44) = 2.50$; $p = 0.080$; BF = 2.61). For all other conditions, differences were present (all $p < 0.05$; all BFs > 13) with more true positive results for t-tests. However, the average differences between the two BDLI methods were negligible, ranging from only 0.03% to 0.35% across conditions. The proportion of false positives was always larger for BLDI with general

linear models than with t-tests (all $p < 0.0001$; all BFs $> 9.6*10^8$), but, again, differences were only small and, across conditions, on average between 0.10% to 1.90%.

We compared the proportion of voxels with correct evidence for h0 across all voxels for which h0 was true (i.e. negatives) between BLDI with t-tests and BLDI with general linear models. Paired t-tests indicated that BLDI with general linear models was superior across all five conditions (all $p < 0.0001$; all BFs $> 3.5*10^{11}$), with about 10-20% more correctly classified voxels across all negatives. Voxels within the simulations regions (i.e. positives) for which Bayesian inference falsely indicated evidence for h0 were rare. While such voxels were occasionally found for BLDI with general linear models, they were almost absent for BLDI with t-tests, or even entirely absent in simulation condition A with a sample size of 50 patients. Because there was little to no variance for this variable in BLDI with t-tests, we refrained from any statistical analysis.

In summary, BLDI methods appeared to be less conservative than VLSM with permutation correction, resulting in more true positives but also more false positives. While BLDI with general linear models was slightly less conservative than BLDI with t-tests, differences were negligibly small. However, BLDI with general linear models was considerably better at finding true evidence for h0.

Figure 5 Results Simulation near here, 2 columns

Figure 6 Results Examples near here, 2 columns

Some aspects of the results deserve a special mention here. First, BFs could reach extremely high values. In the condition with the largest sample size, a few simulations could reach maximum BFs of more than 10^{100}. On the other hand, the smallest BFs were ~0.13, i.e. no BF ever indicated more than moderate evidence for h0. Given this asymmetry, we performed additional tests on random data to investigate how small BFs can become with our given sample sizes. We report these tests in the supplementary. In short, even when h0 was true per definition, BFs indicated only moderate evidence for h0 at best. The ability to collect evidence for h0 was worse with small samples and uneven groups, and, with optimal conditions, i.e. equal groups and a sample size of 300, the minimum BFs were ~0.13 and almost the same as in the simulations.

Second, the performance of all methods, both with Bayesian and null hypothesis significance testing, varied greatly across the 45 simulations (see supplementary tables 1-3). For example, while a method may be completely immune to false alarms in one simulation (i.e. for one specific simulation region), it may suffer massively in the next.

Third, the performance of BLDI with general linear models stood out with a remarkably good performance in simulation condition B which included only small lesions. The overall area of voxels that was tested was comparatively small, which is not surprising given the small number of lesions affecting each voxel (see Figure 1C). However, for voxels that were included in the analysis, Bayesian inference could convincingly differentiate between h0 and h1, often with only very thin borders of voxels for which no evidence was found (see Figure 6 for examples).

Fourth, aside from the analyses with the largest sample size of 300 patients, all analyses mapped only parts of the left brain hemisphere. Large areas, often between one- to two-thirds of the left hemisphere, were lesioned in less than five patients and were thus never statistically tested. Hence, besides areas for which the data provided no evidence in favour of any hypothesis in BLDI (i.e. voxels with a $1/3 < BF < 3$), there was also no evidence in other areas because these areas simply could not be tested due to little to no lesions in this areas.

3.2. In silico validation - posthoc re-analysis with lesion size control

Bayes factor mapping was able to gather evidence for h0 in some areas of the brain – a major advantage over VLSM which does not allow any conclusions of this kind. However, compared to VLSM with permutation correction, we found a pronounced overshoot of evidence for h1 with many false positives, which would limit the use of BLDI in deficits with large effect sizes or large samples. Importantly, BLDI also gathered the most correct evidence for h0 when being used in large samples. As explained in the introduction, an overshoot of lesion-deficit associations is a general problem in lesion-deficit inference, including VLSM with null hypothesis significance testing. This problem affects VLSM to different degrees across multiple comparison correction strategies (Sperber and Karnath, 2017; Mirman et al., 2018; Pustina et al., 2018), and the permutation approach appears to provide very decent protection against false positives (see also Pustina et al., 2018). Lesion size control has been suggested as a possible counter-strategy for analyses with a pronounced overshoot of

associations (Sperber, 2022). With a posthoc analysis, we aimed to evaluate if lesion size control can also improve the performance of BLDI in such situations. We reanalysed the data of simulation condition A with 100 patients with BLDI including lesion size control. We again used general linear models and now compared, in each voxel, a baseline model including the intercept and lesion size to a model that additionally included the binary voxel status. In other words, we tested if the voxel status together with lesion size significantly better explains the deficit than only lesion size.

The results of the post hoc analyses are shown in Figure 7, and detailed results are reported in supplementary table 4. The rate of false evidence for h1 in BLDI was strongly improved by lesion size control (mean (uncontrolled) = 19.5%; mean (controlled) = 11.2%; $t(44) = 5.03$; $p < 0.0001$; BF = 2198). However, visual inspection of example maps (Figure 7) revealed inconsistencies in the performance of lesion size control across the brain. As intended with lesion size control, the areas of evidence for h1 often had smaller peak values and less overshoot of associations beyond the simulation region. However, at the same time, new areas of evidence for h1 appeared in other areas, and the pattern of evidence was altered across the entire brain. This observation was in line with a numerical decrease of correct evidence for h0 after lesion size control (mean (uncontrolled) = 27.6%; mean (controlled) = 23.8%) which, however, was not significant ($t(44) = 1.25$; $p = 0.22$; BF = 0.336). In summary, the voxel-wise consideration of lesion size appeared to combine wanted and unwanted effects that, overall, did hardly improve the quality of the brain maps. Especially the modifications of the brain map in areas that, without lesion size control, indicated no evidence or evidence for h0, were detrimental to the value of BLDI.

Considering these limitations of standard lesion size control, we introduced a modified approach termed *adaptive lesion size control*. This algorithm applied a voxel-wise control for lesion size only in voxels with a Bayes factor > 3 without control, i.e. voxels for which evidence for h1 was found. In voxels for which the uncontrolled analyses did not find evidence in favour of h1, lesion size control was not applied. With this modification of lesion size control, BLDI produced less false evidence for h1 (mean = 6.5%) than uncontrolled BLDI ($t(44) = 7.07$; $p < 0.0001$; BF = $1.4*10^7$), but also less than BLDI with standard lesion size control ($t(44) = 9.16$; $p < 0.0001$; BF = $9.8*10^8$). However, it still created more false evidence for h1 than

VLSM with permutation correction ($t(44) = 8.21$; $p < 0.0001$; $BF = 5.2*10^7$). Yet, we found adaptive lesion size control in BLDI to also improve the collection of correct evidence in favour of h0 (mean = 35.1%) compared to uncontrolled BLDI (mean = 27.6%; $t(44) = 5.53$; $p < 0.0001$; $BF = 1.0*10^4$). In summary, adaptive lesion size control improved BLDI both by a reduction of false positive evidence for h1 and an increase of true positive evidence for h0.

Figure 7 Results ls control near here, 2 columns

3.3. Real-world application of Bayesian lesion-deficit inference

In the analysis of real-world deficits, we tested features with lesion damage in at least 4 patients, which were 214586 voxels. We investigated two deficits for which we assumed varying clinical variance. We expected constructive ability to have high clinical variance and therefore a rather high statistical power, and verbal fluency to have low clinical variance and little statistical power. We referenced results to cortical brain areas with the Automatic Anatomical Labelling atlas (Tzourio-Mazoyer et al., 2002).

The frequentist mapping of verbal fluency found only very minimal and spatially restricted results. The 224 voxels (equivalent to a volume of 0.224cm³) with significant lesion-deficit associations (Figure 8A) were spread across three clusters in different right hemispheric brain areas, with the largest cluster in the rolandic operculum. BLDI implicated ~21,400 voxels with at least moderate evidence for h1 and ~57,000 voxels with evidence for h0 (Figure 8B). Voxels with evidence for h1 peaked in the rolandic operculum, and some clusters were found in the putamen and superior frontal regions. While most neuropsychological studies on language deficits focussed on the left hemisphere, some also investigated the right hemisphere with heterogeneous results (e.g. Stuss et al., 1998; Riello et al., 2021; Biesbroeck et al., 2021). While some did not implicate right hemispheric correlates at all (Riello et al., 2021; Biesbroeck et al., 2021), the present results in superior frontal areas are in line with the findings by Stuss and colleagues. Interestingly, BLDI could provide evidence that the relevance of inferior temporal regions (Riello et al., 2021; Biesbroeck et al., 2021) is not mirrored in the right hemisphere.

Figure 8 Results verbal fluency near here, 2 columns

The frequentist mapping of constructive ability implicated ~29,500 voxels with significant lesion-deficit associations (Figure 9A). These were spread across several grey matter regions including the superior temporal gyrus, rolandic operculum, post- and precentral gyrus, insula, and supramarginal gyrus. A putative pronounced overshoot of associations was found by BLDI (Figure 9B), with a peak BF of ~210,000 and ~150,000 voxels with evidence in favour of h1, and only ~9,000 voxels with evidence for h0. While the voxel-wise map did not allow any precise interpretation by assessing Bayes factors >3, the nuances within the Bayes factor map, i.e. a differentiation between moderate, strong, or at least very strong evidence, highlighted several larger clusters with very strong evidence in the areas where the frequentist analysis found significant results. Given the assumed strong overshoot of lesion-deficit associations, we re-analysed the voxel-wise data by BLDI with adaptive lesion size control. The main cluster of the ~3,900 voxels with evidence for h1 was located in inferior temporal regions and the temporal pole. In the areas implicated in the frequentist analysis, only a few scattered tiny clusters remained.

Figure 9 Results construction ability near here, 2 columns

In summary, the analysis of real-world deficits supported the findings of the in silico study. BLDI was found to be a statistically much more liberal method than frequentist inference. But, still, the foci of the results were qualitatively very similar. Further, the performance of both methods varied across analyses. BLDI appeared to be much more informative when mapping a deficit with small clinical variance (verbal fluency), and frequentist inference when mapping a deficit with large clinical variance and strong effect size (constructive ability).

3.4. Computational performance of Bayes factor mapping

The Bayesian analyses that we used are computationally far more demanding than their non-Bayesian counterparts. The computation time for a single Bayes factor map surpassed the time required for frequentist mapping with null hypothesis significance testing, even though the latter repeated the entire analyse 1000 times for the permutation-derived multiple comparison correction. Still, with an average modern

home computer with an AMD Ryzen 5 3600 with 3.59GHz, the computation time for a single Bayes factor map was manageable. In the in silico validation, about 750.000 voxels were tested in our uni-hemispheric stroke sample at an imaging resolution of 1x1x1 mm^3. Depending on the sample size, this required about 0.5-3 hours of non-parallelised computation time. The inclusion of covariates, such as lesion size, increased this time as multiple models had to be computed for each voxel.

4 Discussion

We implemented different approaches to Bayesian lesion-deficit inference (BLDI) by Bayes factor mapping and compared them to the most common frequentist lesion-deficit approach in an in silico experiment and the mapping of two real-world deficits, verbal fluency and constructive ability. The performance of BLDI, as well as the performance of frequentist lesion-deficit inference, varied across situations. Our results suggest that BLDI is a valuable addition to the method portfolio of lesion-deficit inference with some specific and exclusive advantages, such as the ability to collect evidence for the null hypothesis. However, it is not superior to established frequentist approaches in all respects.

4.1. The advantages of going Bayesian in lesion-deficit inference

The first and likely foremost advantage of BLDI is its ability to gather evidence for the null hypothesis h0. While the ability of BLDI to find such evidence varies across several parameters, large areas with Bayes factors in favour of h0 were found in many analyses in the current study. This leads to new insights not provided by any other approach for lesion-deficit inference and which could significantly provide critical evidence for neuroanatomical theories. Consider, for example, the discussion of the anatomy of spatial neglect that went on for a long time. What shall we conclude from studies that locate the neural correlates of spatial neglect in the parietal cortex (Mort et al., 2003), the temporal cortex (Karnath et al., 2004), or the frontal cortex (Commiteri et al., 2007)? From a frequentist statistical perspective, these results are not mutually exclusive. This is a situation where BLDI could shine and clarify the conclusions. Are we to conclude that the neural correlate of spatial neglect is the parietal cortex but not the temporal cortex, or is it the parietal cortex while the sample provides no information about a possible role of the temporal cortex? With the help of Bayesian statistics, such questions could have been clarified and today's unifying

brain-network theories for spatial neglect might have emerged much earlier.

Besides the evidence for h0, BLDI also provides transparent information on the lack of evidence. While the value of this information for creating and modifying scientific theories is made clear by the previous example of spatial neglect, there is another benefit. The usual method pipeline using frequentist lesion-deficit inference provides no information on how well a sample is suited to map a cognitive function onto the brain. Post hoc mapping of statistical power in voxel-based lesion-symptom mapping is in principle possible (Kimberg et al., 2007). However, it requires an additional topographical analysis that is only provided by some analysis tools and therefore a rare sight in the literature. On the other hand, BLDI provides such information right away within the Bayes factor map. With large areas of Bayes factors close to 1, i.e. Bayes factors that do not provide evidence for either h1 or h0, it becomes obvious that a study lacked statistical power. Further, BLDI can raise awareness of the blind spots of typical lesion-deficit studies. For example, stroke lesions in medial brain areas are rare. Hence, previous stroke studies might have missed finding evidence for the potential contribution of medial areas to spatial neglect (Herbet and Duffaut, 2022). BLDI provides topographical information on this lack of evidence and, thereby, does not only inform us about a general lack of statistical power but also about a lack of statistical power in specific regions, which might be typical for specific lesion aetiologies.

The in silico study part suggested a surprisingly good performance of BLDI in samples with small lesions. Commonly, the size of lesions in lesion-deficit inference comes with a trade-off: small lesions usually provide little overlap (see Figure 1C) and therefore lead to largely uneven groups with low statistical power (compare to Kimberg et al., 2007). On the other hand, small lesions are less likely to damage multiple areas in unison and therefore counter the association problem (see the introduction and Figure 1A). Hence, some authors even suggested primarily relying on small lesions in lesion-deficit inference (Price et al., 2017). As BLDI transparently highlights potential limitations of statistical power and showed such a good performance in our in silico study, it appears to be a good first choice in data sets with small lesions.

Contrary to p-values, Bayes factors provide a meaningful continuous measure for the strength of evidence. This allows a qualitative interpretation of nuances within a statistical brain map to some degree. Using our analysis of constructive ability as an

example (Figure 9), the peaks of the (uncontrolled) Bayes factor map could have been interpreted as the most likely centres of potential neural correlates or, if transparently reported and justified with the association problem, an interpretation could focus on areas with larger Bayes factors only. Likewise, nuances in the statistical maps can be used to compare the degree of evidence between brain areas or to evaluate statistical tendencies and trends.

4.2. The peculiarities and limitations of Bayesian lesion-deficit inference

In general, BLDI appears to be a liberal statistical method in the collection of evidence for the alternative hypothesis h1. In all conditions, BLDI created on average more false-positive evidence for h1 but also more true evidence for h1 than the frequentist counterpart with family-wise error correction. In hindsight, this is no surprise. The conventions for the interpretation of BFs are popular in, e.g., psychology or economics, where studies might aim to explore more subtle effects. In neuropsychology, effect sizes can be massively larger. For example, imagine the effect size in object naming between patients without any aphasic disturbances – which will be perfect most of the time – and patients with a naming deficit. This is not per se a limitation of BLDI, but an aspect that users of BLDI should be aware of. The in silico study found high variance in the performance of BLDI (and frequentist lesion-deficit inference as well) across simulations. In situations with only weak lesion-deficit associations, a liberal statistical test might be at an advantage, as seen in the mapping of verbal fluency. With stronger lesion-deficit associations, a liberal statistical test might overestimate the extent of neural correlates. Hence, it appears that BLDI suffers more from the association problem (as described in the introduction and Figure 1A) than frequentist inference with family-wise error correction. This interpretation is further supported by the results of simulation condition B, in which only small lesions were included in the sample. Small lesions are less likely to damage many brain areas in unison, and therefore they minimise the association problem. In this setting with only a minor impact of the association problem, BLDI provided surprisingly good results both in the collection of evidence for h1 and h0. Of note, the overshoot of evidence for h1 beyond the neural correlate of a function is no false positive evidence in the statistical sense. These associations truly exist in the data. Thus, it appears that the family-wise correction for multiple comparisons in frequentist statistics is very conservative, which can be advantageous in lesion-deficit

inference when the association problem strongly generates causally spurious associations.

Several points are debated as potential general limitations of Bayesian null hypothesis testing (Tendeiro and Kiers, 2019; van Ravenzwaaij and Wagenmakers, 2022). Among those points, one became highly apparent in our study: the collection of evidence for h0 and h1 was asymmetric. While the evidence for h1 accumulated quickly into extreme ranges, evidence for h0 was almost always moderate at best. As seen in the supplementary analyses, sample and group sizes imposed a limit on how small a Bayes factor in favour of h0 can become. With the sample sizes common to lesion-deficit inference, it is unlikely to obtain strong evidence for h0. Given that frequentist statistics were never even able to provide any evidence for h0 at all, we agree with van Ravenzwaaij and Wagenmakers (2022) that the asymmetry in the collection of evidence does not constitute a limitation of Bayesian inference. However, one should be aware that the collection of evidence for h0 has a hard limit imposed by sample size.

The Bayesian approach to lesion-deficit inference that we presented in this work follows the framework of statistical parametric mapping (Friston et al., 1994) with mass-univariate testing. In mass-univariate testing, every single feature of our independent data – such as the status of each voxel or structural connection in the brain – is independently tested, which can result in the execution of many thousand statistical tests. However, the independent statistical assessment of brain features does not represent how the brain is organised by interacting brain regions within networks (Mah et al., 2014; Zhang et al., 2014; Toba et al., 2020). Multivariate, high-dimensional inference methods were proposed as an alternative that is better suited to map the brain (Mah et al., 2014; Zhang et al., 2014; Pustina et al., 2018). It is still debated to what degree multivariate inference solves the limitations of mass-univariate testing (Sperber, 2020; Ivanova et al., 2021). In any case, it appears that multivariate inference is superior in many situations, which include the mapping of deficits that originate from lesions to multiple regions or different areas of large-scale networks (Zhang et al., 2014; Pustina et al., 2018; Ivanova et al., 2021). Hence, multivariate lesion-deficit inference has advantages over BLDI by mass-univariate Bayes factor mapping. However, the advantages of BLDI are specific to this method and, to our knowledge, are not provided by any current multivariate inference method. Therefore, we believe that the different advantages of the two approaches should be

weighed on a situation-by-situation basis and that it can even be a reasonable choice to complement the advantages of both approaches within a multiverse analysis.

4.3. Practical suggestions for the application of Bayesian lesion-deficit inference

A very important conclusion from the current study should be that no perfect method for lesion-deficit inference exists. The performance of all methods varied across situations and conditions and, in general, both frequentist and Bayesian methods were affected by the spatial distribution of lesions across the brain and the varying coverage of lesions across regions. The current study, as well as several previous studies on the validity of lesion-deficit inference methods (Zhang et al., 2014; Pustina et al., 2018; Ivanova et al., 2021; Sperber, 2022) highlight that method performance varies depending on the situation and data parameters, and that for some data sets, each method has difficulty capturing the neural correlates of a deficit. Research questions and a priori assumptions on the organisation of a neural correlate factor into the choice of a method. But still, many data parameters are difficult to predict, such as lesion distribution or variance of a deficit, and could justify transparently reported post hoc changes in a methodological pipeline.

Bayesian hypothesis testing introduces a new category of non-evidence compared to frequentist lesion-deficit inference. Bayes factors around 1 indicate that the data do not provide considerable evidence for any of the hypotheses h1 and h0. Clandestinely, many more brain features exist in lesion-deficit inference for which also no evidence is present, and they even already existed in the frequentist approach. As shown in Figure 2B and Figure 3B, only in parts of the brain sufficient variance exists in the imaging variable to perform a meaningful statistical test at all. This terra incognita of lesion-deficit inference and the ‘no evidence’ category of Bayesian hypothesis testing are conceptually very similar. Both could be considered in the generation and evaluation of theories on brain anatomy in the same way. An intuitive strategy to transparently differentiate both categories in data visualisation is based on the fact that Bayes factors are always larger than 0, and, as in the current study, zeros within the statistical map inform the reader about any features that were not tested.

Our study suggests that BLDI is well-suited for lesion data samples with low statistical power or small lesions. Both situations are problematic in frequentist approaches. When statistical power is low, the first advantage of BLDI is that it is a relatively liberal statistical approach in the collection of evidence for h1. The second

advantage is that, even if the first advantage should not help in a given situation, BLDI is unprecedently transparent with the statistical power limitation. When lesions are small, we found that, within the few features that can be meaningfully tested, BLDI was very precise in delineating the border of a neural correlate often with good confidence to decide for either h1 or h0. On the other hand, BLDI can be too liberal in the detection of evidence for h1 when statistical power is high. Hence, changes to conventions for the interpretation of BFs could be considered. The interpretation of BFs > 3 as at least moderate evidence (Wagenmakers et al., 2011; 2018) and similar conventions are not universal, and, in general, existing conventions were not generated with the possibly huge effects of neuropsychology in mind. With knowledge about the data distribution of a specific neuropsychological data set in mind, one might decide on different conventions, e.g., to consider only evidence with BFs larger than 10, 30, or 100, corresponding to strong, very strong, or extreme evidence in the conventions suggested by Wagenmakers and colleagues (2018). More generally, the comparability of Bayesian and frequentist statistics not only in lesion-deficit inference but in all kinds of imaging analyses remains a challenge. Existing frequentist analysis designs offer established conventions for statistical data interpretation, including common α-levels and multiple comparison correction strategies, but a straightforward conversion into Bayes factors is not possible. Hence, future studies are required to optimise the adaption of statistical imaging analysis with Bayesian statistics while maintaining comparability with previous works.

4.3.1. Lesion size control

The flexibility of Bayesian general linear models allows for the inclusion of covariates such as lesion size. Lesion size could potentially counteract the association problem (Sperber and Karnath, 2017) which, as the present study suggests, is a greater burden for BLDI than for frequentist approaches. However, lesion size is no true confound and, therefore, lesion size control is not generally valid and guaranteed to improve the precision of results (Sperber, 2022; see also Wysocki et al., 2022 for more general information on statistical control). Accordingly, the impact of standard lesion size control provided conflicting results. To improve its impact, we modified the standard approach for lesion size control. *Adaptive* lesion size control that only controlled regions for which evidence for h1 existed in the first place outperformed standard lesion size control, and both the ability of BLDI to collect evidence for h1

and h0 were improved over uncontrolled BLDI. However, the performance of this lesion size control still varied and, in the mapping of constructive ability, it provided little additional insights. As previously explained by Sperber (2022), we believe a flexible and transparent approach to be most suited. Given the many unpredictable parameters in lesion studies, it is difficult to estimate a priori the strength of lesion-deficit associations. Therefore, we believe that it can well be justified to decide on the application of lesion size control post hoc, whenever topographical results appear to suffer from an overshoot of associations. Controlled results should be interpreted with caution though, as the control for lesion size could overshadow true lesion-deficit associations. This could well have been the case in our study when mapping constructive ability, where large areas were controlled away. Lesion size control should be most viable in samples with high statistical power, i.e. samples with large effect or sample sizes. Although we examined the effects of lesion size control only for the baseline sample of 100 patients to save computational resources, we assume that its effects would have been even more beneficial for the largest sample size (as in Sperber, 2022).

4.4. Generalisation of Bayes factor mapping to other brain imaging data and test designs

The wide success of univariate statistical parametric mapping is likely rooted in its simplicity and flexibility. It can provide an analysis framework for almost any kind of brain mapping in diseased brains, with the potential to include different statistical tests suited for all kinds of data and designs. Bayesian inference by Bayes factor mapping follows the same conceptual framework and thereby adapts these advantages of statistical parametric mapping, as well as some of its limitations as discussed in the previous section. In the current study, we utilised Bayesian t-tests and general linear models, which are both suited for the mapping of continuous variables. Analysis of non-continuous variables, such as binary or ordinal measures of deficits, require, e.g., logistic regression, ordinal regression, or contingency table tests. The Bayesian variants of these tests are available for example in the R BayesFactor package (Morey et al., 2018) or the R BFpack (Mulder et al., 2021) and can be applied in the same way as Bayesian t-tests. These packages also include more complex multivariate statistical models that allow the consideration of secondary variables, for example within multiple regression, MANOVAs, and general linear models. They open up the

opportunity to account for confounding cognitive or behavioural deficits, lesion size, or other clinical variables.

Bayes factor mapping is not limited to voxel-wise lesion data. The design with Bayesian general linear models can be transferred, for example, to topographical lesion-network maps (as in Boes et al., 2015), topographical white matter alterations (as in Umarova et al., 2014) or structural disconnection maps (as in Foulon et al., 2018, Griffis et al., 2021). Of course, Bayes factor mapping is not limited to topographic data. In the supplementary analyses, we applied it to parcel-to-parcel disconnection matrices and, likewise, it can be used with any other imaging feature or clinical variable. The few previous studies that used lesion-deficit inference with Bayes factors are good examples of the method's flexibility. For example, they investigated the neural correlates of two different apraxia measures with a Bayesian analysis of variance (Achilles et al., 2017) and the association of voxel-wise disconnection severity with post-stroke fatigue (Ulrichsen et al., 2021).

4.5. Limitations

A major limitation of in silico studies in lesion mapping is the limited external validity. In silico studies can highlight general aspects, limitations, and tendencies in lesion-deficit inference, but they will be unable to exactly represent all the complex parameters of clinical data, such as the impact of interindividual differences, secondary variables and co-morbidity, or imperfect spatial normalisation. We included random noise in the in silico experiment to make the data more representative of such parameters, but, still, lesion-deficit associations were sometimes extremely high, and higher than the associations found in real-world data. Hence, the in silico study experiment might have been over-confident in the effect sizes in neuropsychological data, and, therefore, the impact of the association problem on BLDI might have been overestimated. In previous works of ourselves and others, we often encountered data for which frequentist VLSM found no significant voxels, even in sample sizes over 100 patients. Therefore, the limitation that BLDI is too liberal might be less of a burden in real samples, as we have also seen in the mapping of verbal fluency in the present study. Further, the degree to which our findings can be transferred to other stroke samples will likely vary depending on factors such as recruitment parameters, acute treatment, or study design. Likewise, we advise caution

to not draw any conclusions on the general adequacy of a method for certain brain regions based on our findings, as our results may also depend on sampling effects.

Data availability and BLDI Software

The study code and extended results including the statistical maps are publicly available at data.mendeley.com/datasets/5ztswgzhvy/2.

We published an R-based toolkit for Bayesian lesion-deficit inference by Bayes factor mapping in lesion and disconnection data at github.com/ChrisSperber/BLDI and data.mendeley.com/datasets/k6dcbkjdcx/1.

Declaration of Competing Interest

There are no conflicts of interest to report.

Funding

This work is funded by the Synapsis Foundation and the Anna Seiler Foundation (Grant number 2019-PI05).

References

Achilles, E.I.S., Weiss, P.H., Fink, G.R., Binder, E., Price, C.J., Hope, T.M.H., 2017. Using multi-level Bayesian lesion-symptom mapping to probe the body-part-specificity of gesture imitation skills. Neuroimage 161, 94–103. https://doi.org/10.1016/j.neuroimage.2017.08.036

Arnoux, A., Toba, M.N., Duering, M., Diouf, M., Daouk, J., Constans, J.M., Puy, L., Barbay, M., Godefroy, O., 2018. Is VLSM a valid tool for determining the functional anatomy of the brain? Usefulness of additional Bayesian network analysis. Neuropsychologia 121, 69–78. https://doi.org/10.1016/j.neuropsychologia.2018.10.003

Aschenbrenner, S., Tucha, O., Lange, K.W., 2000. Regensburger Wortflüssigkeits-Test: RWT. Hogrefe, Verlag für Psychologie.

Bates, E., Wilson, S.M., Saygin, A.P., Dick, F., Sereno, M.I., Knight, R.T., Dronkers, N.F., 2003. Voxel-based lesion-symptom mapping. Nat. Neurosci. 6, 448–50. https://doi.org/10.1038/nn1050

Behroozmand, R., Bonilha, L., Rorden, C., Hickok, G., Fridriksson, J., 2022. Neural correlates of impaired vocal feedback control in post-stroke aphasia. Neuroimage 250, 118938. https://doi.org/10.1016/j.neuroimage.2022.118938

Bhaskar, S., Stanwell, P., Cordato, D., Attia, J., Levi, C., 2018. Reperfusion therapy in acute ischemic stroke: dawn of a new era? BMC Neurol. 18, 8. https://doi.org/10.1186/s12883-017-1007-y

Biesbroek, J.M., Lim, J.S., Weaver, N.A., Arikan, G., Kang, Y., Kim, B.J., Kuijf, H.J., Postma, A., Lee, B.C., Lee, K.J., Yu, K.H., Bae, H.J., Biessels, G.J., 2021. Anatomy of phonemic and semantic fluency: A lesion and disconnectome study in 1231 stroke patients. Cortex 143, 148–163. https://doi.org/10.1016/j.cortex.2021.06.019

Bonkhoff, A.K., Lim, J.-S., Bae, H.-J., Weaver, N.A., Kuijf, H.J., Biesbroek, J.M., Rost, N.S., Bzdok, D., 2021. Generative lesion pattern decomposition of cognitive impairment after stroke. Brain Commun. 3. https://doi.org/10.1093/braincomms/fcab110

Bonkhoff, A.K., Bretzner, M., Hong, S., Schirmer, M.D., Cohen, A., Regenhardt, R.W., Donahue, K.L., Nardin, M.J., Dalca, A. V, Giese, A., Etherton, M.R., Hancock, B.L., Mocking, S.J.T., McIntosh, E.C., Attia, J., Benavente, O.R., Bevan, S., Cole, J.W., Donatti, A., Griessenauer, C.J., Heitsch, L., Holmegaard, L., Jood, K., Jimenez-Conde, J., Kittner, S.J., Lemmens, R., Levi, C.R., McDonough, C.W., Meschia, J.F., Phuah, C.-L., Rolfs, A., Ropele, S., Rosand, J., Roquer, J., Rundek, T., Sacco, R.L., Schmidt, R., Sharma, P., Slowik, A., Söderholm, M., Sousa, A., Stanne, T.M., Strbian, D., Tatlisumak, T., Thijs, V., Vagal, A., Wasselius, J., Woo, D., Zand, R., McArdle, P.F., Worrall, B.B., Jern, C., Lindgren, A.G., Maguire, J., Fox, M.D., Bzdok, D., Wu, O., Rost, N.S., Bonkhoff, A.K., Bretzner, M., Hong, S., Schirmer, M.D., Cohen, A., Regenhardt, R.W., Donahue, K.L., Nardin, M.J., Dalca, A. V, Giese, A., Etherton, M.R., Hancock, B.L., Mocking, S.J.T., McIntosh, E.C., Attia, J., Benavente, O.R., Bevan, S., Cole, J.W., Donatti, A., Griessenauer, C.J., Heitsch, L., Holmegaard, L., Jood, K., Jimenez-Conde, J., Kittner, S.J., Lemmens, R., Levi, C.R., McDonough, C.W., Meschia, J.F., Phuah, C.-L., Rolfs, A., Ropele, S., Rosand, J., Roquer, J., Rundek, T., Sacco, R.L., Schmidt, R., Sharma, P., Slowik, A., Söderholm, M., Sousa, A., Stanne, T.M., Strbian, D., Tatlisumak, T., Thijs, V., Vagal, A., Wasselius, J., Woo, D., Zand, R., McArdle, P.F., Worrall, B.B., Jern, C., Lindgren, A.G., Maguire, J., Fox, M.D., Bzdok, D., Wu, O., Rost, N.S., 2022. Sex-specific lesion pattern of functional outcomes after stroke. Brain Commun. 4, 1–12. https://doi.org/10.1093/braincomms/fcac020

Boes, A.D., Prasad, S., Liu, H., Liu, Q., Pascual-Leone, A., Caviness, V.S., Fox, M.D., 2015. Network localization of neurological symptoms from focal brain lesions. Brain 138, 3061–75. https://doi.org/10.1093/brain/awv228

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–76. https://doi.org/10.1038/nrn3475

Chen, R., Herskovits, E.H., 2010. Voxel-based Bayesian lesion-symptom mapping. Neuroimage 49, 597–602. https://doi.org/10.1016/j.neuroimage.2009.07.061

Chen, R., Hillis, A.E., Pawlak, M., Herskovits, E.H., 2008. Voxelwise Bayesian lesion-deficit analysis. Neuroimage 40, 1633–1642. https://doi.org/10.1016/j.neuroimage.2008.01.014

Committeri, G., Pitzalis, S., Galati, G., Patria, F., Pelle, G., Sabatini, U., Castriota-Scanderbeg, A., Piccardi, L., Guariglia, C., Pizzamiglio, L., 2007. Neural bases of personal and extrapersonal neglect in humans. Brain 130, 431–41. https://doi.org/10.1093/brain/awl265

Dienes, Z., 2014. Using Bayes to get the most out of non-significant results. Front. Psychol. 5, 1–17. https://doi.org/10.3389/fpsyg.2014.00781

Duering, M., Gesierich, B., Seiler, S., Pirpamer, L., Gonik, M., Hofer, E., Jouvent, E., Duchesnay, E., Chabriat, H., Ropele, S., Schmidt, R., Dichgans, M., 2014. Strategic white matter tracts for processing speed deficits in age-related small vessel disease. Neurology 82, 1946–1950. https://doi.org/10.1212/WNL.0000000000000475

Foulon, C., Cerliani, L., Kinkingnéhun, S., Levy, R., Rosso, C., Urbanski, M., Volle, E., Thiebaut de Schotten, M., 2018. Advanced lesion symptom mapping analyses and implementation as BCBtoolkit. Gigascience 7, 1–17. https://doi.org/10.1093/gigascience/giy004

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. 2, 189–210. https://doi.org/10.1002/hbm.460020402

Gallistel, C.R., 2009. The Importance of Proving the Null. Psychol. Rev. 116, 439–453. https://doi.org/10.1037/a0015251

Gelman, A., Carlin, J., 2014. Beyond Power Calculations. Perspect. Psychol. Sci. 9, 641–651. https://doi.org/10.1177/1745691614551642

Griffis, J.C., Metcalf, N. V., Corbetta, M., Shulman, G.L., 2021. Lesion Quantification Toolkit: A MATLAB software tool for estimating grey matter damage and white matter disconnections in patients with focal brain lesions. NeuroImage Clin. 30, 102639. https://doi.org/10.1016/j.nicl.2021.102639

Herbet, G., Duffau, H., 2022. Contribution of the medial eye field network to the voluntary deployment of visuospatial attention. Nat. Commun. 13, 328. https://doi.org/10.1038/s41467-022-28030-3

Ivanova, M. V., Herron, T.J., Dronkers, N.F., Baldo, J. V., 2021. An empirical comparison of univariate versus multivariate methods for the analysis of brain–behavior mapping. Hum. Brain Mapp. 42, 1070–1101. https://doi.org/10.1002/hbm.25278

Karnath, H.-O., Sperber, C., Rorden, C., 2018. Mapping human brain lesions and their functional consequences. Neuroimage 165. https://doi.org/10.1016/j.neuroimage.2017.10.028

Karnath, H.-O., Fruhmann Berger, M., Küker, W., Rorden, C., 2004. The anatomy of spatial neglect based on voxelwise statistical analysis: a study of 140 patients. Cereb. Cortex 14, 1164–72. https://doi.org/10.1093/cercor/bhh076

Karnath, H.-O., Rennig, J., 2017. Investigating structure and function in the healthy human brain: validity of acute versus chronic lesion-symptom mapping. Brain Struct. Funct. 222, 2059–2070. https://doi.org/10.1007/s00429-016-1325-7

Keysers, C., Gazzola, V., Wagenmakers, E.J., 2020. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. Nat. Neurosci. 23, 788–799. https://doi.org/10.1038/s41593-020-0660-4

Kimberg, D.Y., Coslett, H.B., Schwartz, M.F., 2007. Power in Voxel-based lesion-symptom mapping. J. Cogn. Neurosci. 19, 1067–80. https://doi.org/10.1162/jocn.2007.19.7.1067

Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O., 2008. Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc. 103, 410–423. https://doi.org/10.1198/016214507000001337

Mah, Y.-H., Husain, M., Rees, G., Nachev, P., 2014. Human brain lesion-deficit inference remapped. Brain 137, 2522–31. https://doi.org/10.1093/brain/awu164

Mirman, D., Landrigan, J.-F., Kokolis, S., Verillo, S., Ferrara, C., Pustina, D., 2018. Corrections for multiple comparisons in voxel-based lesion-symptom mapping. Neuropsychologia 115, 112–123. https://doi.org/10.1016/j.neuropsychologia.2017.08.025

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A.(2018). Bayes Factor (Version 0.9.12-4.4) [computer software]. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Morris, J.C., Heyman, A., Mohs, R.C., Hughes, J.P., van Belle, G., Fillenbaum, G., Mellits, E.D., Clark, C., 1989. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology 39, 1159–65. https://doi.org/10.1212/wnl.39.9.1159

Mulder, J., Williams, D.R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink, M., Menke, J., van Aert, R., Fox, J.-P., Hoijtink, H., Rosseel, Y., Wagenmakers, E.-J., van Lissa, C., 2021. BFpack : Flexible Bayes Factor Testing of Scientific Theories in R. J. Stat. Softw. 100. https://doi.org/10.18637/jss.v100.i18

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum. Brain Mapp. 15, 1–25. https://doi.org/10.1002/hbm.1058

Price, C.J., Hope, T.M., Seghier, M.L., 2017. Ten problems and solutions when predicting individual outcome from lesion site after stroke. Neuroimage 145, 200–208. https://doi.org/10.1016/j.neuroimage.2016.08.006

Pustina, D., Avants, B., Faseyitan, O.K., Medaglia, J.D., Coslett, H.B., 2018. Improved accuracy of lesion to symptom mapping with multivariate sparse

canonical correlations. Neuropsychologia 115, 154–166. https://doi.org/10.1016/j.neuropsychologia.2017.08.027

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Riello, M., Frangakis, C.E., Ficek, B., Webster, K.T., Desmond, J.E., Faria, A. V, Hillis, A.E., Tsapkini, K., 2021. Neural Correlates of Letter and Semantic Fluency in Primary Progressive Aphasia. Brain Sci. 12, 1. https://doi.org/10.3390/brainsci12010001

Rorden, C., Fridriksson, J., Karnath, H.-O., 2009. An evaluation of traditional and novel tools for lesion behavior mapping. Neuroimage 44, 1355–62. https://doi.org/10.1016/j.neuroimage.2008.09.031

Rorden, C., Karnath, H.O., 2004. Using human brain lesions to infer function: A relic from a past era in the fMRI age? Nat. Rev. Neurosci. 5, 812–819. https://doi.org/10.1038/nrn1521

Rorden, C., Karnath, H.O., Bonilha, L., 2007. Improving lesion-symptom mapping. J. Cogn. Neurosci. 19, 1081–1088. https://doi.org/10.1162/jocn.2007.19.7.1081

Rouder, J.N., Morey, R.D., 2012. Default Bayes Factors for Model Selection in Regression. Multivariate Behav. Res. 47, 877–903. https://doi.org/10.1080/00273171.2012.734737

Sperber, C., Karnath, H.-O., 2017. Impact of correction factors in human brain lesion-behavior inference. Hum. Brain Mapp. 38. https://doi.org/10.1002/hbm.23490

Sperber, C., Wiesen, D., Karnath, H.-O., 2019. An empirical evaluation of multivariate lesion behaviour mapping using support vector regression. Hum. Brain Mapp. 40. https://doi.org/10.1002/hbm.24476

Sperber, C., Griffis, J., Kasties, V., 2022. Indirect structural disconnection-symptom mapping. Brain Struct. Funct. 227, 3129–3144. https://doi.org/10.1007/s00429-022-02559-x

Sperber, C., 2020. Rethinking causality and data complexity in brain lesion-behaviour inference and its implications for lesion-behaviour modelling. Cortex 126. https://doi.org/10.1016/j.cortex.2020.01.004

Sperber, C., 2022. The strange role of brain lesion size in cognitive neuropsychology. Cortex 146, 216–226. https://doi.org/10.1016/j.cortex.2021.11.005

Stuss, D.T., Alexander, M.P., Hamer, L., Palumbo, C., Dempster, R., Binns, M., Levine, B., Izukawa, D., 1998. The effects of focal anterior and posterior brain lesions on verbal fluency. J. Int. Neuropsychol. Soc. 4, 265–278. https://doi.org/10.1017/s1355617798002653

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Toba, M.N., Godefroy, O., Rushmore, R.J., Zavaglia, M., Maatoug, R., Hilgetag, C.C., Valero-Cabré, A., 2020. Revisiting 'brain modes' in a new computational era: approaches for the characterization of brain-behavioural associations. Brain 143, 1088–1098. https://doi.org/10.1093/brain/awz343

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289. https://doi.org/10.1006/nimg.2001.0978

Ulrichsen, K.M., Kolskår, K.K., Richard, G., Alnæs, D., Dørum, E.S., Sanders, A.M., Tornås, S., Sánchez, J.M., Engvig, A., Ihle-Hansen, H., de Schotten, M.T., Nordvik, J.E., Westlye, L.T., 2021. Structural brain disconnectivity mapping of post-stroke fatigue. NeuroImage Clin. 30. https://doi.org/10.1016/j.nicl.2021.102635

Umarova, R.M., Reisert, M., Beier, T.U., Kiselev, V.G., Klöppel, S., Kaller, C.P., Glauche, V., Mader, I., Beume, L., Hennig, J., Weiller, C., 2014. Attention-network specific alterations of structural connectivity in the undamaged white matter in acute neglect. Hum. Brain Mapp. 35, 4678–4692. https://doi.org/10.1002/hbm.22503

Umarova, R.M., Saur, D., Kaller, C.P., Vry, M.-S., Glauche, V., Mader, I., Hennig, J., Weiller, C., 2011. Acute visual neglect and extinction: distinct functional state of the visuospatial attention system. Brain 134, 3310–3325. https://doi.org/10.1093/brain/awr220

van Doorn, J., Ly, A., Marsman, M., Wagenmakers, E.-J., 2020. Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. J. Appl. Stat. 47, 2984–3006. https://doi.org/10.1080/02664763.2019.1709053

van Ravenzwaaij, D., & Wagenmakers, E.-J. (2022). Advantages masquerading as "issues" in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). Psychological Methods, 27(3), 451–465. https://doi.org/10.1037/met0000415

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H.L.J., 2011. Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). J. Pers. Soc. Psychol. 100, 426–432. https://doi.org/10.1037/a0022790

Wagenmakers, E.J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q.F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J.N., Morey, R.D., 2018. Bayesian inference for psychology. Part II:

Example applications with JASP. Psychon. Bull. Rev. 25, 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wetzels, R., Grasman, R.P.P.P., Wagenmakers, E.J., 2012. A default Bayesian hypothesis test for ANOVA designs. Am. Stat. 66, 104–111. https://doi.org/10.1080/00031305.2012.695956

Wetzels, R., Wagenmakers, E.J., 2012. A default Bayesian hypothesis test for correlations and partial correlations. Psychon. Bull. Rev. 19, 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Wysocki, A.C., Lawson, K.M., Rhemtulla, M., 2022. Statistical Control Requires Causal Justification. Adv. Methods Pract. Psychol. Sci. 5, 251524592210958. https://doi.org/10.1177/25152459221095823

Zhang, Y., Kimberg, D.Y., Coslett, H.B., Schwartz, M.F., Wang, Z., 2014. Multivariate lesion-symptom mapping using support vector regression. Hum. Brain Mapp. 35, 5861–5876. https://doi.org/10.1002/hbm.22590

Zhao, Y., Halai, A.D., Lambon Ralph, M.A., 2020. Evaluating the granularity and statistical structure of lesions and behaviour in post-stroke aphasia. Brain Commun. 2, 1–14. https://doi.org/10.1093/braincomms/fcaa062

Figure 1: Problems in lesion-deficit inference(A) Illustration of the association problem in lesion-deficit inference. For this example, assume that a lesion to the STG is the sole cause of a cognitive deficit. Due to the typical anatomy of stroke, the lesion load to the STG is associated with the lesion load in many other regions, such as the insula, the inferior parietal lobule, and the frontal inferior operculum. Therefore, the lesion load in these regions might as well be associated with the deficit. These associations are no direct causal relationships. However, common statistical tests are unable to differentiate between direct causal relationships and associations. The shown correlations were computed in the present sample of 300 patients. **(B)** Percentage of patients with a lesion in a voxel in a sample of 300 stroke patients. The statistical power of voxel-based lesion-symptom mapping is highest in voxels with equal groups, i.e., at a percentage of 50%, when 150 patients have a lesion and 150 have no lesion. **(C)** Percentage of patients with a lesion in a voxel in a sample of 150 stroke patients with small lesions. The sample was from the 300 patients included in Figure 1B after a median split of lesion size, i.e., only the 150 patients with the smallest lesions were included. **(D)** Illustration of the statistical limitations resulting from the partial injury of a neural correlate. See also Sperber et al. (2019) for a more detailed illustration of this issue. The data underlying Figures 1A-C are the 300 lesion maps that we used in the in silico study part.

Figure 2: Lesion topography in the sample of 300 patients(A) Lesion overlap topography of all 300 patients included in the in silico study part. **(B)** The same lesion overlap thresholded for voxels lesioned in at least five patients, i.e. voxels included in the lesion-deficit analyses in the condition with the largest sample of 300 patients.

Even with this comparatively very large sample, some medial areas in the left hemisphere were not analysed at all.

Figure 3: Lesion topography in the sample of 137 patients
(A) Lesion overlap topography of all 137 patients included in the study part on real-world deficits. The numbers above the slices indicate the z-coordinate in MNI space. **(B)** The same lesion overlap thresholded for voxels lesioned in at least four patients, i.e. voxels included in the lesion-deficit analyses.

Figure 4: Concept of the in silico study part

Figure 5: Results of the in silico study part
Average performance of frequentist voxel-based lesion-symptom mapping (VLSM) and Bayesian lesion-deficit inference (BLDI) with either Bayesian t-tests or general linear models across all simulation conditions with 45 simulation runs each. The underlying data were transformed into values representing the proportion of relevant voxels in % across all possible positive voxels (i.e. the simulation region) respectively all negative voxels (i.e. all voxels outside of the simulation region that were damaged at least in five patients in the total sample as shown in Figure 2B). Note that the bars on the right represent far more voxels than the left ones. Condition A compared different sample sizes (50-100-300), condition B looked at small lesions only with a sample of 100 patients, and condition C included a second simulation region also with 100 patients. Detailed results including standard deviation and range are reported in supplementary tables 1-3.

Figure 6: Results of the in silico study part – example maps
Results of the in silico study for three simulations comparing voxel-based lesion-symptom mapping (VLSM) with null hypothesis significance testing and Bayesian lesion-deficit inference (BLDI) by Bayesian generals linear models (GLMs). The columns (A50 to C100) represent the different simulation conditions. The Bayes factors were binned into categories for clarity of visualisation; in practice, this is not advisable, as continuous Bayes factors convey meaningful information about the strength of evidence. Areas with moderate or strong evidence in favour of h1 are rare compared to areas with very strong evidence for h1 (BF > 30); they usually only frame the areas where there is evidence of h0 and are barely visible in the figure. Additional topographical results are provided in the online materials.

Figure 7: The influence of lesion size control
Results of post hoc analyses of lesion-deficit inference in simulation condition A with examples. In addition to Figures 5 and 6, the results of Bayesian lesion-deficit inference (BLDI) either with standard lesion size control or adaptive lesion size

control are shown. In two out of the 45 simulations, BLDI with lesion size control now indicated strong evidence for h0 (BF < 1/10) in at least some voxels. As such was still a rare occurrence, we do not report this category separately.

Figure 8: Results of frequentist and Bayesian lesion deficit inference for verbal fluency

(A) Results of frequentist lesion-deficit inference with null hypothesis significance testing and family-wise error correction via maximum statistic permutation at $\alpha = 0.05$. Significant voxels were scarce and are highlighted in blue. **(B)** Results of Bayesian lesion-deficit inference.

Figure 9: Results of frequentist and Bayesian lesion deficit inference for constructive ability
Results of **(A)** frequentist and **(B)** Bayesian lesion deficit inference. **(C)** Voxel-wise Bayesian lesion deficit inference with adaptive lesion size control.

Credit_author_statement

Christoph Sperber: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft. **Laura Gallucci**: Investigation, Data curation, Writing - Review & Editing. **Stefan Smaczny**: Conceptualization, Software Validation, Writing - Review & Editing. **Roza Umarova:** Conceptualization, Resources, Data curation, Writing- Reviewing and Editing, Funding acquisition

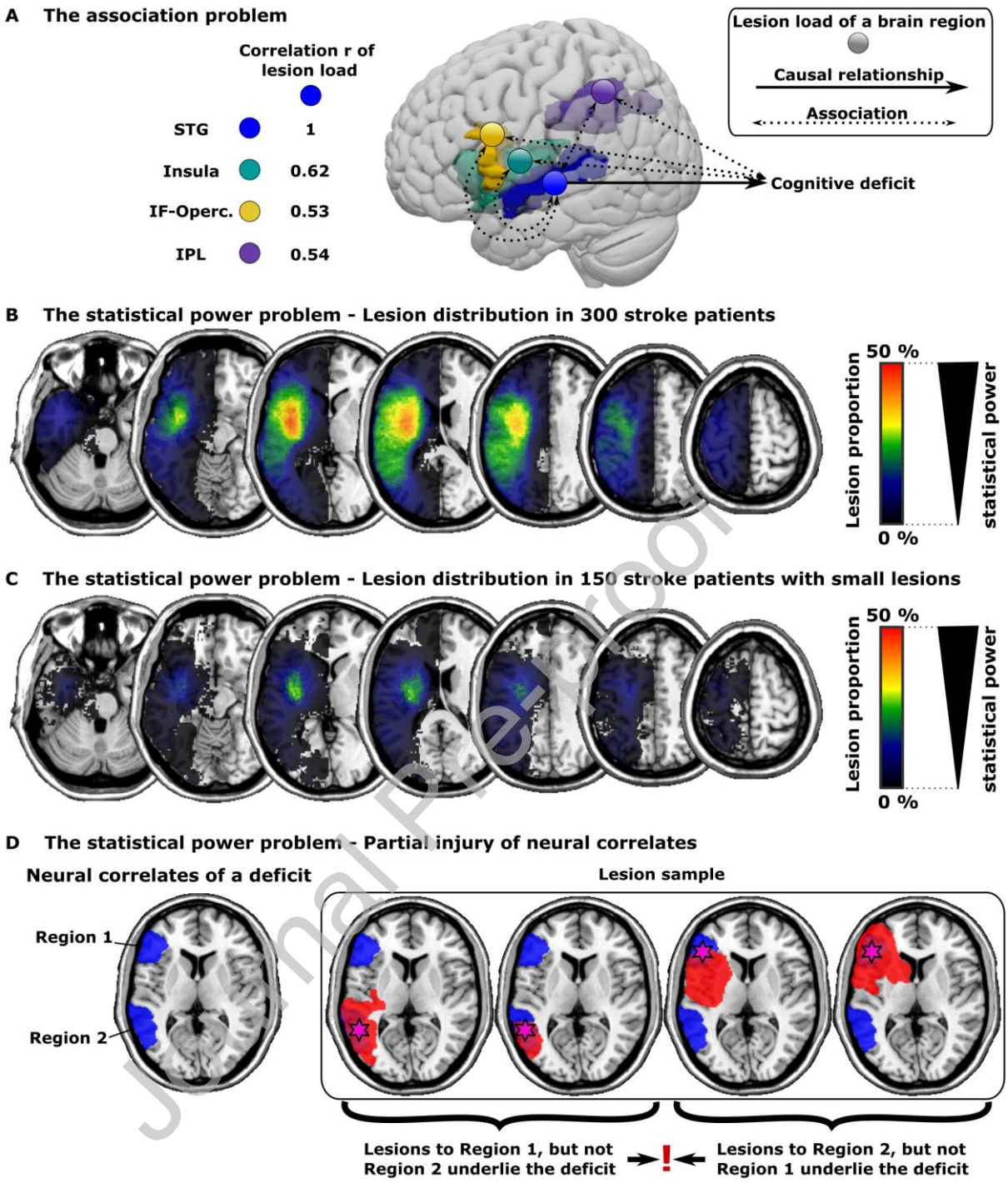


Fig 1

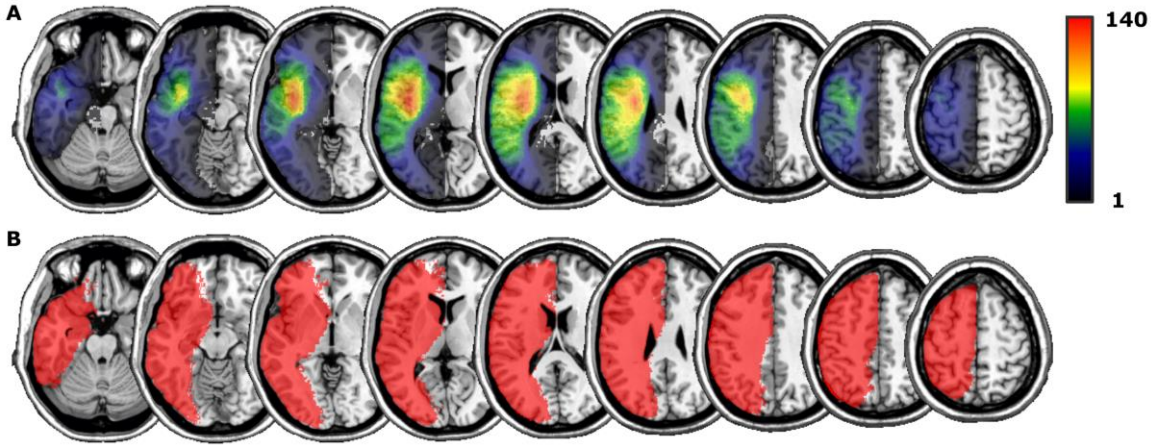


Fig 2

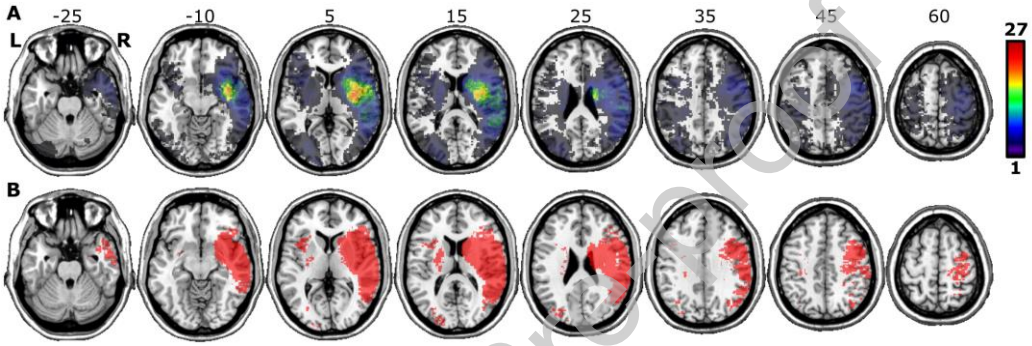


Fig 3

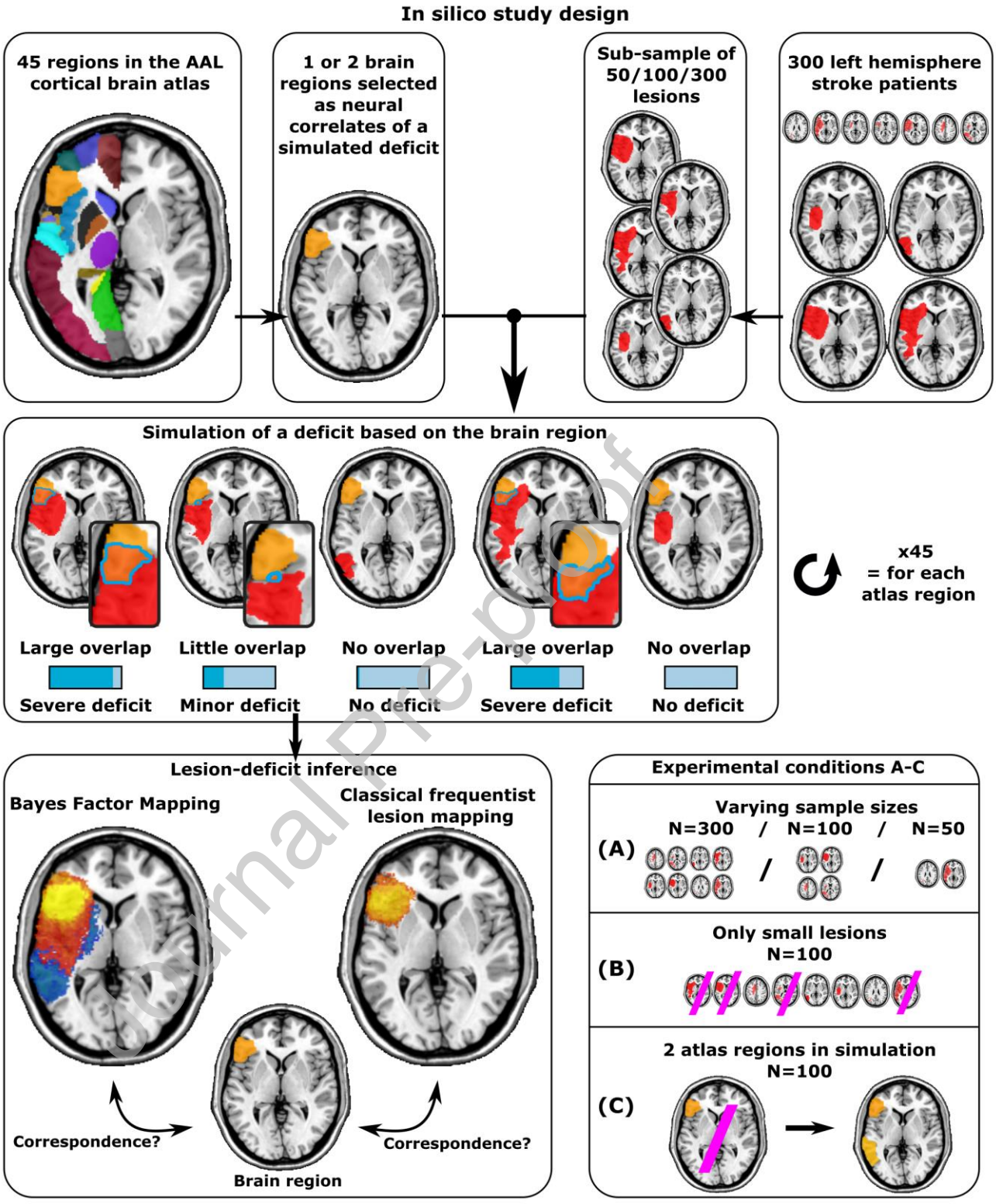

In silico study design
45 regions in the AAL cortical brain atlas
1 or 2 brain regions selected as neural correlates of a simulated deficit
Sub-sample of 50/100/300 lesions
300 left hemisphere stroke patients
Simulation of a deficit based on the brain region
Large overlap
Little overlap
No overlap
Large overlap
No overlap
Severe deficit
Minor deficit
No deficit
Severe deficit
No deficit
x45
= for each atlas region
Lesion-deficit inference
Bayes Factor Mapping
Classical frequentist lesion mapping
Correspondence?
Correspondence?
Brain region
Experimental conditions A-C
Varying sample sizes
N=300 / N=100 / N=50
(A)
Only small lesions
N=100
(B)
2 atlas regions in simulation
N=100
(C)

Fig 4

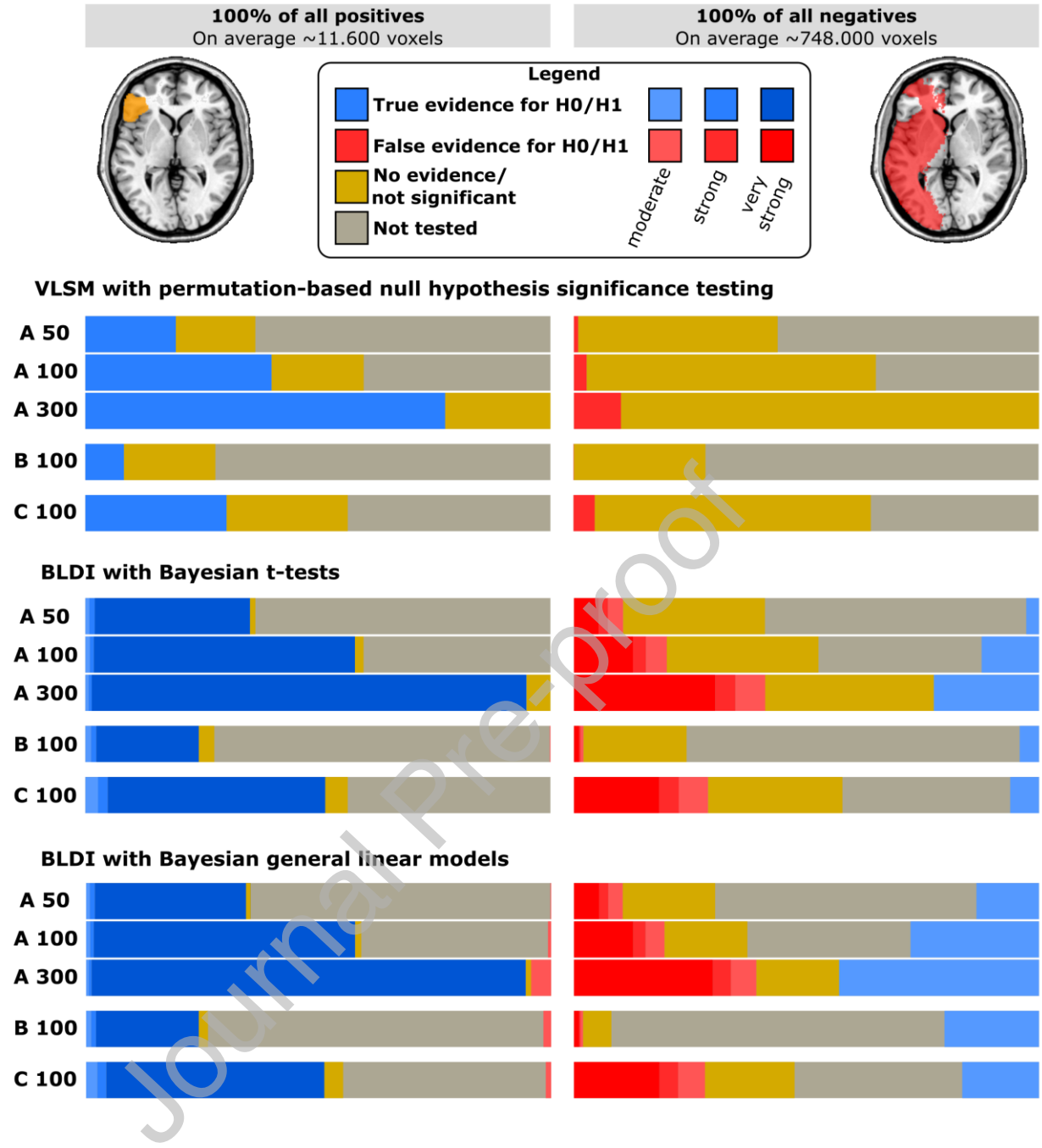


Fig 5

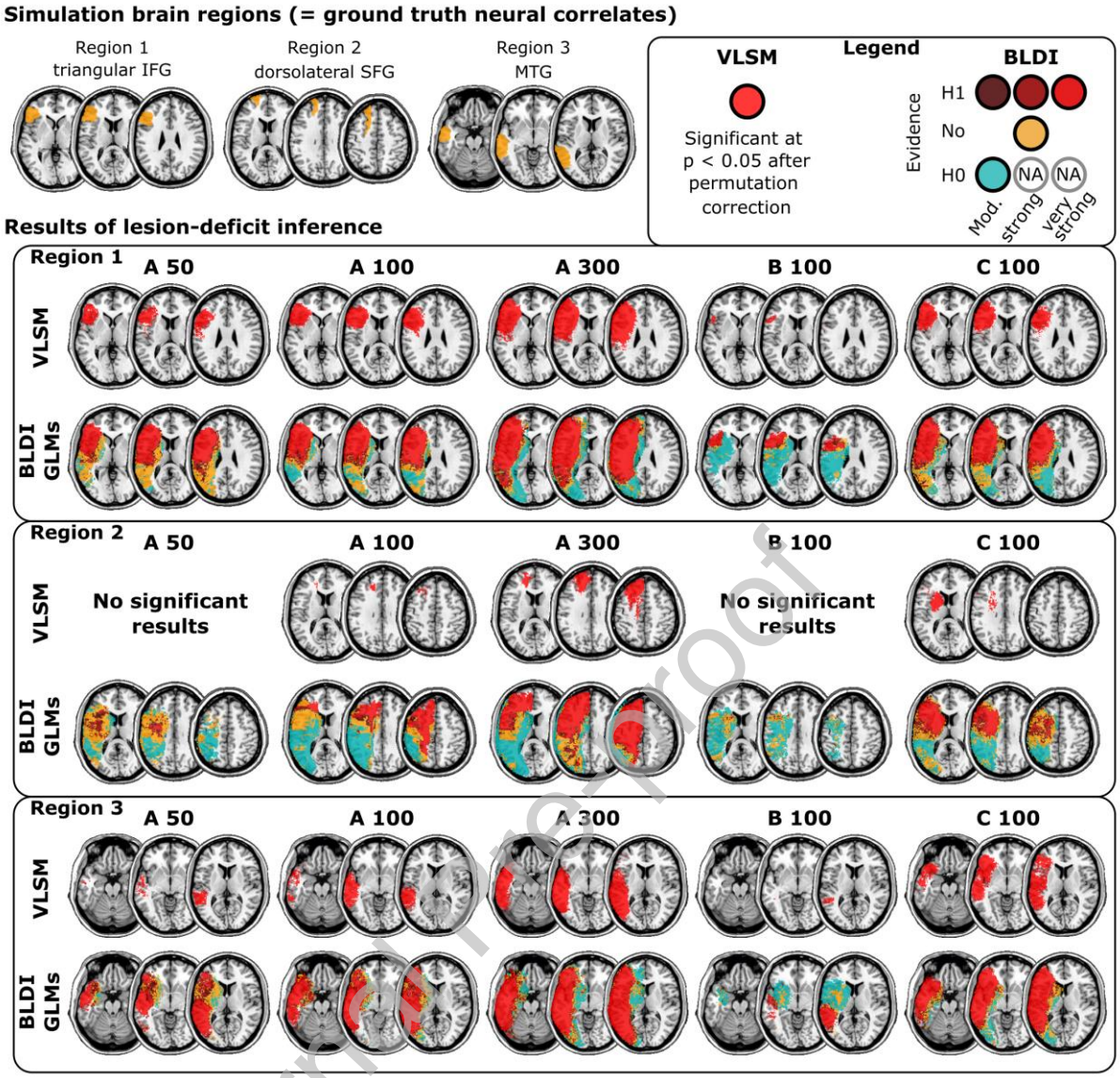


Fig 6

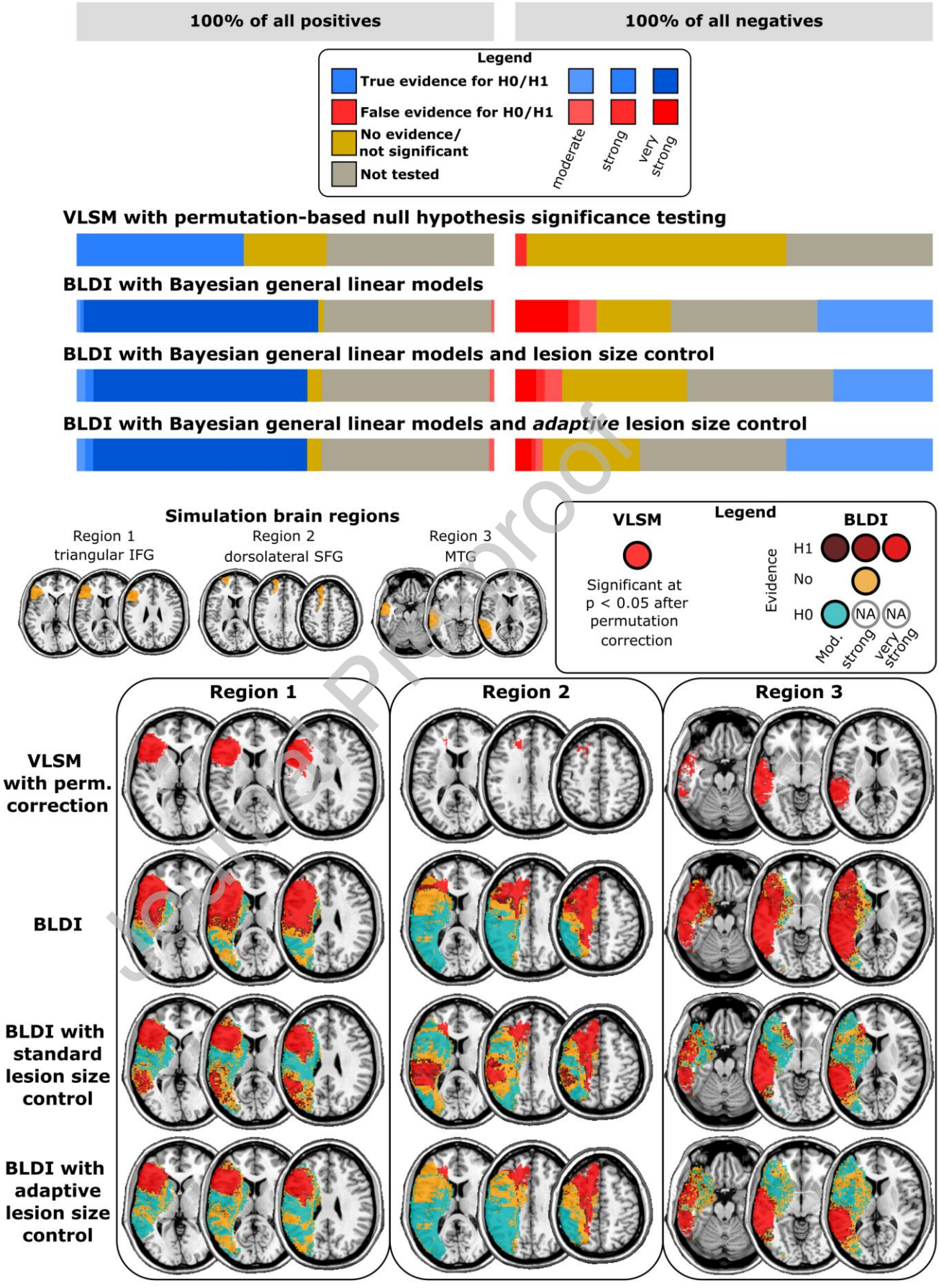

100% of all positives
100% of all negatives
Legend
True evidence for H0/H1
False evidence for H0/H1
No evidence/ not significant
Not tested
moderate
strong
very strong
VLSM with permutation-based null hypothesis significance testing
BLDI with Bayesian general linear models
BLDI with Bayesian general linear models and lesion size control
BLDI with Bayesian general linear models and adaptive lesion size control
Simulation brain regions
Region 1 triangular IFG
Region 2 dorsolateral SFG
Region 3 MTG
Legend
VLSM
Significant at p < 0.05 after permutation correction
BLDI
Evidence
H1
No
H0
NA
NA
Mod.
strong
very strong
Region 1
Region 2
Region 3
VLSM with perm. correction
BLDI
BLDI with standard lesion size control
BLDI with adaptive lesion size control

Fig 7

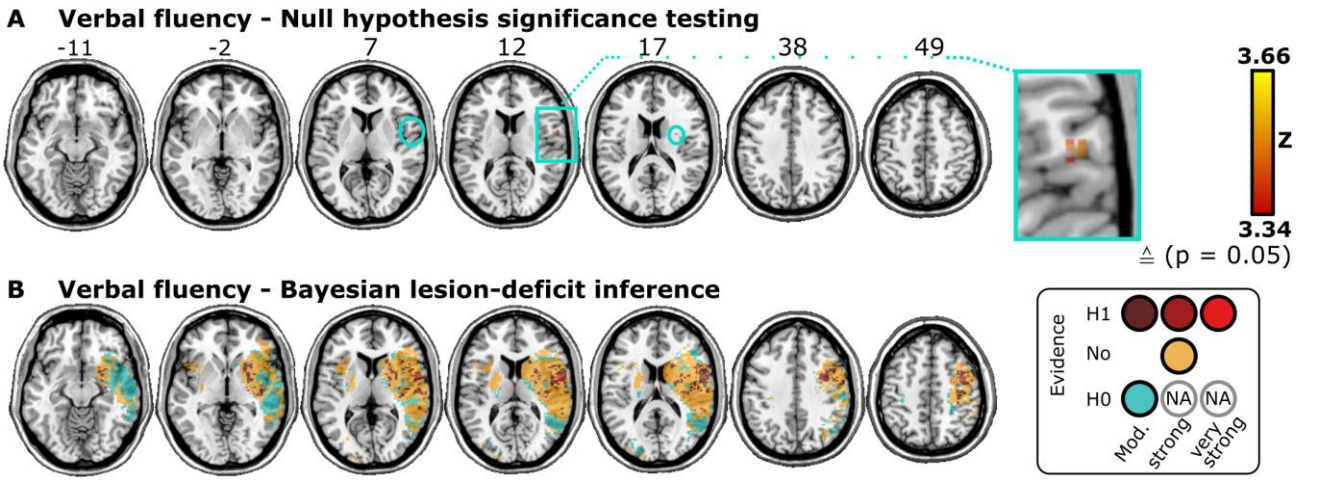


Fig 8

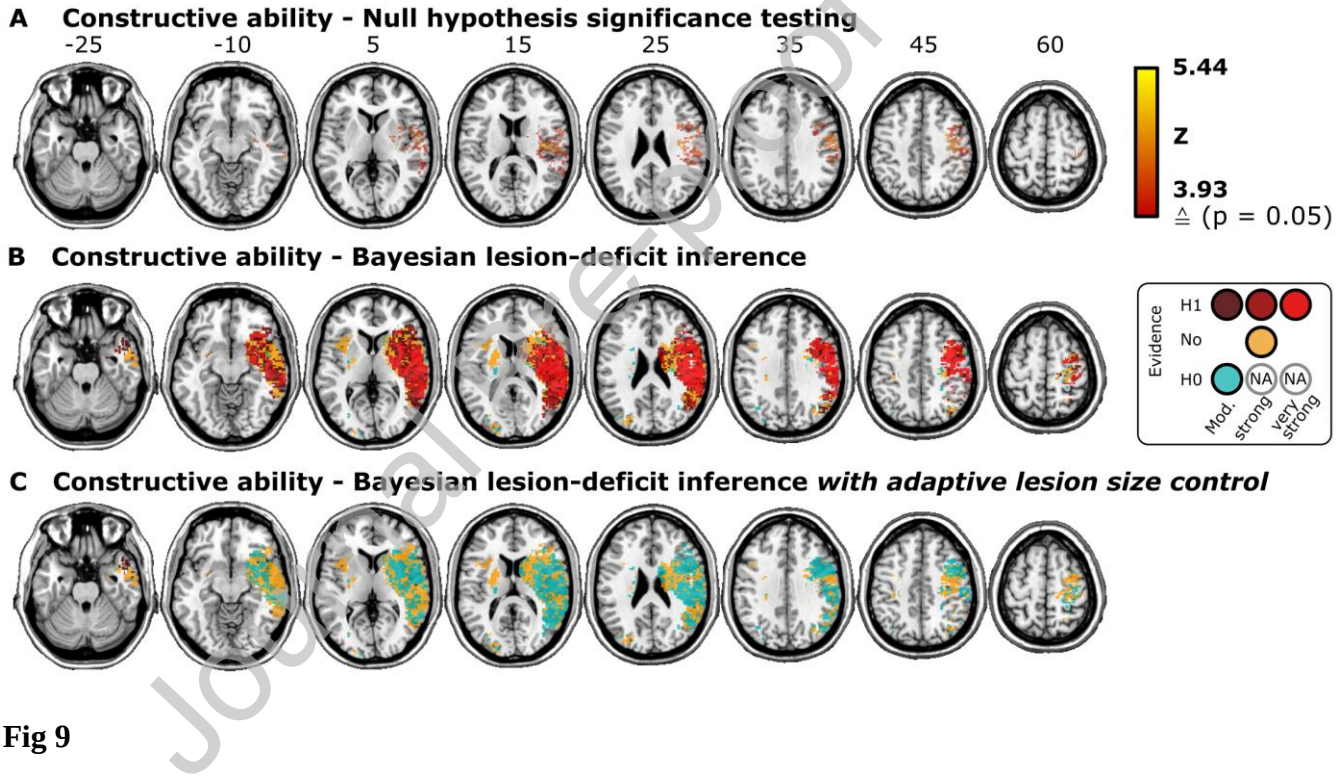
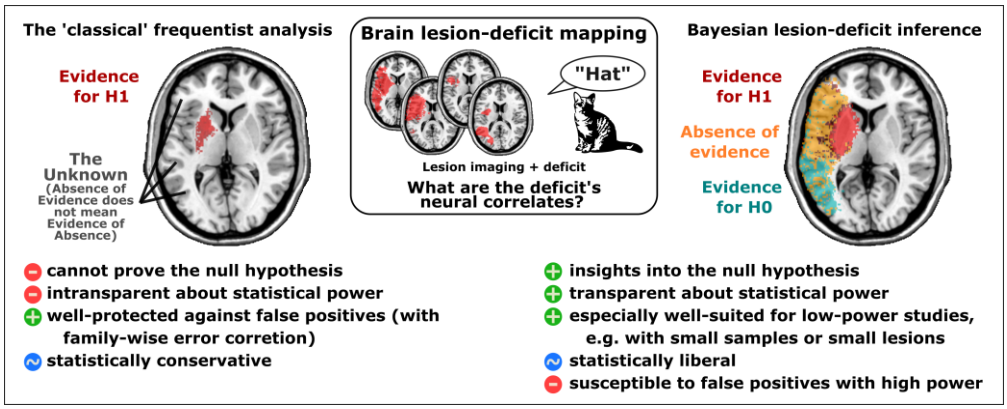


Fig 9



Graphical abstract