Unsupervised mapping of causal relations between brain and behavior

Iman A. Wahle[1,2], Ralph Adolphs[1,2], Daniel Tranel[3], Aaron Boes[3,4,5,6], Frederick Eberhardt[2]

[1] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA

[2] Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA

[3] Department of Neurology, University of Iowa Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA

[4] Department of Pediatrics, University of Iowa Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA

[5] Department of Psychiatry, University of Iowa Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA

[6] Department of Psychological and Brain Sciences, University of Iowa, Iowa City, Iowa, USA

Correspondence to fde@caltech.edu

Abstract

Studies of patients with focal brain lesions provided the historical foundation for cognitive neuroscience, but how to identify a precise mapping between specific brain regions and the cognitive variables affected remains unclear. The challenge lies both in identifying anatomical regions wherein lesions have a shared causal effect, as well as in the precise delineation of the behavioral outcome. Currently, either the relevant brain region or the dimensionality of the behavior being mapped are pre-specified by the investigators rather than both being informed by optimal brain-behavior relationships. Here we apply a novel data-driven causal aggregation algorithm, Causal Feature Learning (CFL) to tackle this challenge in 520 individuals with focal brain lesions. CFL simultaneously constructs macro-level summaries of the spatial distribution of brain lesions and the itemized responses on psychometric tests to optimally characterize the causal brain-behavior relationships. Focusing on the domains of language, visuospatial ability, and depression, we recapitulate established findings, provide new and more precise anatomical results, and present an aggregation of item-wise data that provides an empirical test of extant behavioral scores and can be used to identify novel, psychologically meaningful factors. Future work could use our approach to construct entirely new psychometric variables that might cut across established categories.

Keywords: lesions, depression, language, visuospatial, causal models, dimensionality

Introduction

Lesion research background

Ever since the seminal work of Broca, Wernicke, and others [1], demonstrably large and often remarkably specific cognitive impairments have been documented following macroscopic but focal brain lesions. Such case studies launched much of modern cognitive neuroscience [2,3] and were the first to suggest that specific neural structures might be most essential for memory [4], aspects of face perception[5,6], conscious vision[7], and complex decision-making [8]. Since the advent of CT and structural MRI, larger samples of subjects and better quantification of lesion location provided considerably more detailed, and more statistically reliable associations between lesion locations and specific cognitive impairments in domains ranging from facial emotion recognition [9] to aspects of language [10], object recognition [11], intelligence [12], and many other domains.

In the absence of controlled experimental interventions, as are common in animal studies, human lesion studies offer a route to investigating the causal relations between brain regions and behavioral outcomes. Their interventional nature, even if accidents of nature rather than experimentally controlled, makes the results less prone to suffering from confounding effects, a concern often raised about similar findings based on large-scale functional neuroimaging studies in healthy subjects [13,14]. However, the inference to a causal conclusion in lesion studies still faces the double challenge of establishing a causal *relation* while simultaneously determining the causal *relata*: at what granularity should brain regions and behavioral outcomes be analyzed to identify the specific brain regions that cause specific behavioral outcomes [15,16]? Identifying such brain regions is crucial to the development of novel treatments, but over-aggregation or mis-aggregation of voxel-level measurements can obscure the true causal relations present in the data (see Figure 1A). Simply put: Given the brain lesion and change in task performance, how can the lesion-to-behavior map best be characterized as a generalizable causal relation?

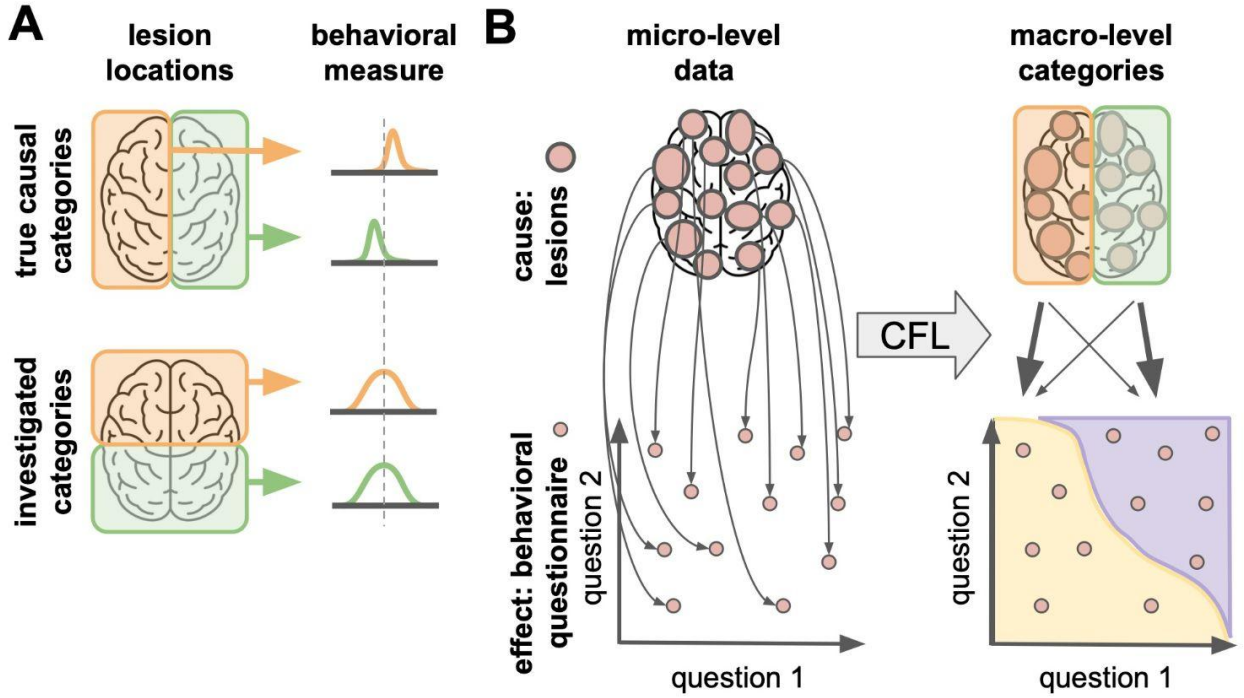


***Figure 1**: Causal feature aggregation. A) Toy example of lesion mis-aggregation. While in the true system, left-hemispheric lesions cause higher behavioral scores than right-hemispheric lesions, looking at behavioral scores corresponding to mis-aggregated lesion categories (anterior versus posterior) obscures the true effects. B) CFL aggregates cause and effect in a completely data-driven way, based upon the actual relations observed in the micro-level data (that is, patients with specific lesions in one location or another, exhibiting specific performance on a task or questionnaire). Macro-level cause categories (orange, green) are distinct in their effects; macro-level effect categories (yellow, purple) are distinct in their likelihoods given the cause.*

Currently, lesion-behavior studies use voxel-wise mappings [17,18] or one of their multivariate extensions[19] to identify which regions of the brain best predict the behavioral outcomes when lesioned. These maps, in general, fit a penalized linear regression from the lesioned voxels to an aggregate quantity (a score) of the behavioral aspects of interest (language ability, visuospatial reasoning, depression, etc.). Depending on the exact method, the regression is subject to different additional constraints: in the case of voxel-wise maps the constraints amount to ensuring that the resulting brain region(s) are locally contiguous and not too small – neighboring voxels should behave similarly, such that regression on a spatially isolated voxel is considered to be noise. In the multivariate extensions, various versions of Canonical Correlation Analysis (CCA) are used. Instead of requiring a contiguity constraint on the identified brain regions, CCA fits an (optionally sparse) regression through a bottleneck of fixed dimensionality that compresses the signal. The outcome is similar to the voxelwise mapping in that locally contiguous brain regions are often identified. In both cases, the methods depend on a host of free parameters (contiguity constraints, bottleneck width, sparsity penalties, etc.) that are chosen by the researcher. Moreover, on the effect side, all measures are set a priori: The

approaches assume that the pre-defined behavioral assessment (generally, a score that summarizes performance across multiple individual items), actually provides a representation of the behavioral impairment at the level of granularity that captures the effects of lesions. Such pre-setting of the outcome dimensions imposes major limitations on one of the most exciting opportunities that lesion studies offer: that an analysis of the effects of lesions might result in dissociations of behavior that we would *not* predict a priori.

Here we present an alternative, more data-driven approach to the identification of the lesion-behavior maps. Causal Feature Learning (CFL), a machine learning method we developed [20–22], is explicitly designed to preserve causal relations in the aggregation from the micro-level measurement data to the macro-level causal representation, without assuming that either the macro-level cause or effect are pre-defined, let alone that they are linearly related (Figure 1B). CFL estimates a non-parametric conditional probability distribution P(Y|X) between the high-dimensional lesion data (X) and the full vector of itemized responses on the cognitive tests (Y). To the extent that the lesions (X) can be viewed as interventions, P(Y|X) represents the causal effect of the lesions (i.e. CFL would be estimating the causal quantity P(Y|do(X))). However, since certain categories of lesions and behavioral patterns may still be associated due to confounding causes, such as particular demographics, we also check whether available demographic measures might confound the effects we see. CFL then aggregates this fine-grained *relation* between the voxel-level descriptions of the lesions and the itemized responses in an unsupervised manner to more optimally identify the causal relationship between specific brain regions and categories of behavioral responses.

We thus obtain a data-driven macro-level description that identifies anatomical features (the lesion location) that reliably correspond to particular response patterns on the behavioral tests (the participants' performance profiles). CFL automatically detects the optimal granularity at which to study the lesion-behavior maps: lesions at possibly distinct locations that have the same effect on behavioral responses will be grouped together, and similarly, cognitive test responses are grouped together if, with respect to the effects of the lesions, they are closely related. Conversely, two lesions may end up in distinct cause categories if their effects on the behavioral measures are different, *even if* they partly overlap in the voxel space.

On the effect side, individuals with similar total test scores may be disaggregated if the lesion data suggests a dissociation of behavior at the level of individual test items. If, on the other hand, established behavioral measures track the causal effect of lesions on behavior well, then CFL can provide an independent confirmation of these scores and suggest brain regions that are candidate causes of the behavior tracked by these measures. In contrast to many existing studies, this approach permits a brain-wide search for the relation between lesions and behavioral outcomes. Inevitably, the search is limited to the lesions for which there exist data (and by the behavioral measures that were obtained), but CFL avoids the need to extract specific (often very small) subsets of subjects to investigate the behavioral effect of a particular category of lesion (such as, e.g., one study by Koenigs et al. [23], which had n=7 patients with lesions in a target region (vmPFC) contrasting with 5 patients with lesions in the dlPFC), or for investigators to define the dimensionality of the behavior to be mapped.

In the following, we validate CFL as a novel method by identifying brain-behavior relationships in a large dataset of participants with focal lesions (distributed across the brain, albeit not homogeneously) and corresponding neuropsychological test scores. First, we evaluate a proof-of-concept analysis to assess whether CFL identifies a well-established association of left-sided lesions with language deficits and right-sided lesions with visuospatial deficits [24–26]. Next, we apply CFL to depressed mood, where the brain-behavior associations from lesion location have been difficult to establish despite several decades of study [27]. It is possible that attempts to infer brain-behavior associations in depression have been challenging, in part, because depression is a syndrome that includes distinct types of symptoms that may individually localize to different anatomical systems. By using CFL with all individual item-level responses of the Beck Depression Inventory (as opposed to a single total BDI score), we can evaluate whether data-driven clustering of individual questionnaire items coupled with data-driven clustering of lesion location can derive more robust brain-behavior associations with depression. We can thereby empirically assess whether the summary BDI score for depression is sufficient to track the lesion-induced behavioral effects measured by the individual questions, or whether an alternative, possibly multi-dimensional, aggregate of the questions would be more informative. In particular, we can determine whether more robust brain-behavior relationships exist with specific BDI items grouped together, which could include recognized item groupings (cognitive versus somatic-affective symptoms) or completely novel groupings. In addition, we compare our findings to results obtained using a standard CCA to demonstrate the superior predictive power of the data-driven brain-behavior relationships identified using CFL.

Participants, behavioral measures, and neuroimaging data

All participants provided written informed consent and all procedures were approved by the Institutional Review Boards of the participating institutions and are in accordance with the Declaration of Helsinki. The patient cohorts for the following two analyses were drawn from two databases.

Cohort 1: The first cohort included 501 individuals from the Patient Registry of the Division of Neuropsychology and Cognitive Neuroscience at the University of Iowa Department of Neurology. The Registry includes patients who suffered brain damage from ischemic stroke, subarachnoid hemorrhage, or neurosurgical resection. Neuroimaging was done in the chronic epoch (>3 months) after lesion onset. Each patient completed a battery of neuropsychological tests, in particular, the Controlled Oral Word Association test (COWA), which evaluates language production, and the Judgment of Line Orientation test (JLO), which evaluates visuospatial discrimination.

Cohort 2: For our analysis of depressed mood, we evaluated a second cohort of 520 individuals derived from two data sources. 325 patients were included from the Patient Registry of the Division of Neuropsychology and Cognitive Neuroscience at the University of Iowa Department of Neurology described above (individuals partially overlapping with the first cohort). In addition, data from 195 individuals were included from the Vietnam Head Injury Study (VHIS) registry, which is composed of American veterans who suffered brain damage mainly from penetrating

head injuries in the Vietnam War. For these patients, neuroimaging was obtained 36-39 years post-injury. For all individuals across both sites, the itemized responses for the Beck Depression Inventory-II (BDI-II) are available. This is a 21-item self-reported questionnaire that evaluates characteristic attitudes and symptoms of depression experienced over the preceding two weeks. Data collection details for both cohorts are included in the supplementary material.

Lesion masks for each patient were generated as described in Bowren et al. (2020). Briefly, lesions were manually traced by trained experts from MRI or CT scans and subsequently transformed into MNI space. The lesion masks from all patients were coregistered to the 1mm MNI152 T1 template brain, and quality checked for accuracy in all cases. Lesion masks for all participants were smoothed from 1mm to 1cm isotropic voxels by setting the value of the coarsened voxel to the proportion of the lesioned voxels within it. COWA, JLO, and BDI scores are sums of their corresponding test item responses. BDI items are measured on a 0-3 scale. Heatmaps of the lesion densities and histograms of the COWA and JLO responses are shown for cohort 1 subjects used in our language and visuospatial analysis in Figure 2. Figure 3 shows the corresponding lesion mask heatmaps and response histograms for the 21 BDI items for cohort 2 subjects used in the depression analysis. In addition, we show the correlation matrix of BDI responses and the correlations of the aggregate somatic-affective and cognitive sub-scores of BDI [28].

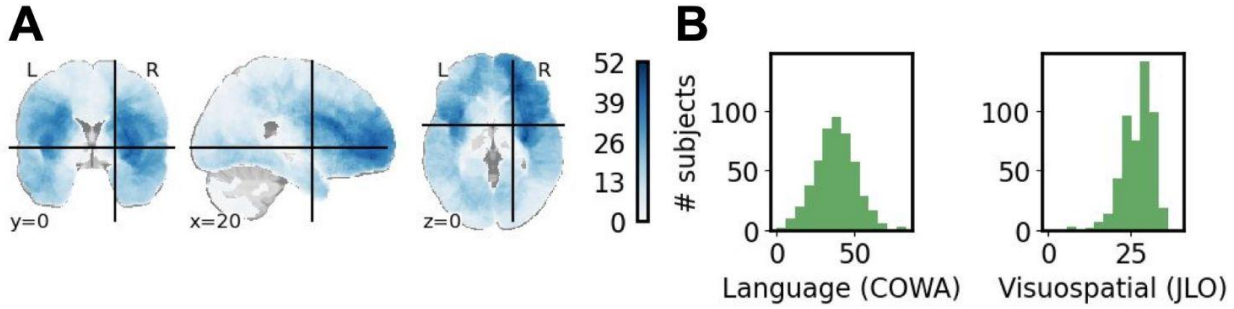


***Figure 2**: 501 individuals (cohort 1) included for the language and visuospatial analysis. A) Heatmap of lesion density across the individuals at central representative slices. Crosshairs indicate the locations of slices in other perspectives. B) Left: Histogram of scores in this cohort on the Controlled Oral Word Association test (COWA). Right: Histogram of scores in this cohort on the Judgment of Line Orientation (JLO) test (the Pearson correlation between the two scores is r = 0.15).*

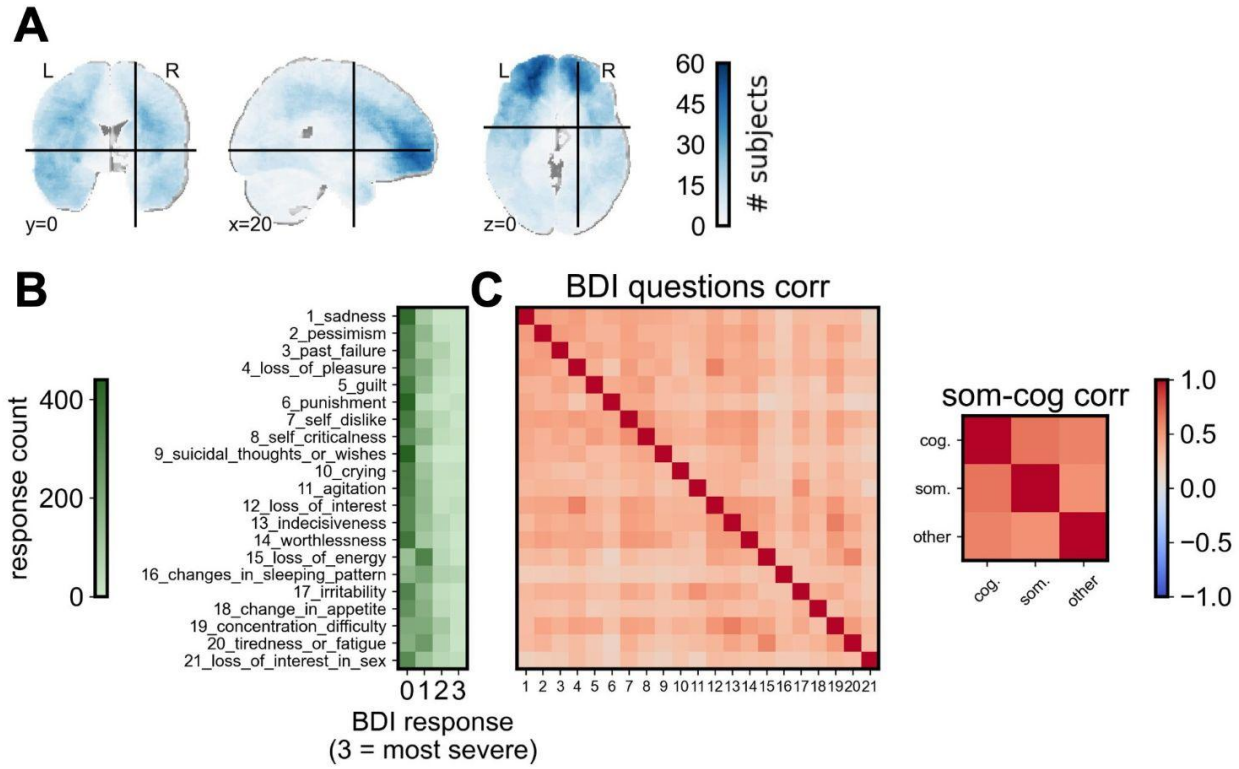


***Figure 3**: 520 individuals (cohort 2) included in the depression analysis. A) Heatmap of lesion density across the individuals at central representative slices. Crosshairs indicate the locations of slices in other perspectives. B) Response distribution for these participants on the 21 items of the BDI questionnaire. C) Pearson correlation for the 21 BDI items (left). Pearson correlation matrix of the cognitive and somatic-affective scores that aggregate the 21 BDI items (right).*

Lesion-behavior mapping

Lesion-behavior maps were learned using Causal Feature Learning (CFL). CFL identifies a mapping between cause variable X (individual lesion masks) and effect variable Y (individual neuropsychological test scores) by 1) estimating the high-dimensional conditional probability of test scores given lesion location P(Y | X) and 2) grouping lesion masks with similar conditional probabilities over the test score outcomes. This method stands in contrast with traditional clustering applications since the clustering is based on the *relation* between lesion masks and test score outcomes, rather than clustering the lesion masks and responses independently and then learning the relation.

In practice, given the comparatively small sample size[1], we configured CFL with a ridge regression that predicts test scores from lesion masks to approximate P(Y | X), and a KMeans clustering model to partition lesion masks based on this learned relation. We chose the ridge regression regularization parameter *alpha* based on out-of-sample prediction error. The number

[1] Originally, we fit a feed-forward neural network that resulted in qualitatively similar findings, but found that with a sample size of only about 500 subjects, our results were not sufficiently robust against different network initializations.

of clusters *K* used by the K-means model was selected by optimizing the recoverability of the conditional distribution from the resulting partition. Model configuration, fitting, and hyperparameter tuning details can be found in the methods section [Methods: Causal Feature Learning (CFL)].

For each analysis, CFL categorized the lesion masks according to the effects the lesions had on neuropsychological test scores. Lesions with similar causal effects are placed in the same category. For the BDI analysis, CFL additionally identified groups of itemized BDI responses that are similarly affected by the distribution of lesion masks. We compared this grouping to that which arises when we solely use the standard summary BDI statistic (sum of responses to questions on the BDI questionnaire) to obtain an empirical evaluation of the suitability of this coarser summary statistic to describe depressive symptoms.

Results

CFL identifies established language and visuospatial brain regions from data

To demonstrate that CFL can recover field-established results in a completely data-driven manner, we first used the method to analyze the relationship between lesions and language and visuospatial performance. We applied the CFL algorithm to cohort 1 to identify distinct categories of lesion masks that are unique in their effects on language and visuospatial assessments.

CFL identifies five categories of lesions with distinct behavioral outcomes (Figure 4, see also Fig S2 for more detail). Although CFL does not enforce any kind of spatial localization or contiguity of lesions within a category, four of five categories do show spatial specificity. Inspection of the location of these lesion categories that CFL identifies with particular behavioral outcomes shows an alignment with past lesion results that are associated with distinctive language and visuospatial scores. In particular, left-hemispheric lesions isolated in categories C1 (and to some extent in C5) correspond to below-average language scores, with respect to average scores for cohort 1. The two categories distinguish between larger left-hemispheric lesions (C1) that are associated with more severe language deficits, and smaller left-hemispheric lesions with less severe language effects (C5), as can be seen in the differences in lesion size distribution in Figure 4A. Category C2, with right hemisphere lesions, corresponds to below-average visuospatial scores. Language performance is not significantly affected in this category. Category C3, with average-sized lesions in the prefrontal cortex, is identified with the surprising effect of above-average language and visuospatial skills [29]. Category C4, which includes the largest proportion of subjects, shows no lesion localization and corresponds to average language and visuospatial scores - lesions that don’t have specific effects on either test.

The unsupervised delineation of left-hemispheric and right-hemispheric lesions and their respective associations with language and visuospatial deficits demonstrates the efficacy of CFL

as a data-driven method for lesion-behavior mapping. The isolation of frontal cortex lesions corresponding to higher language and visuospatial scores demonstrates the advantage of using a data-driven approach to identify all lesion-behavior relationships present in a dataset at the appropriate level of granularity. If we had instead taken a hypothesis-driven approach by comparing all left and right hemisphere lesions, 1) finer-grained distinctions between lesions within each hemisphere would have been obscured and 2) the distinct (and opposite!) effect of frontal cortex lesions, which may occur in the left or right hemisphere but is not associated with language and visuospatial deficits, would not have been identified.

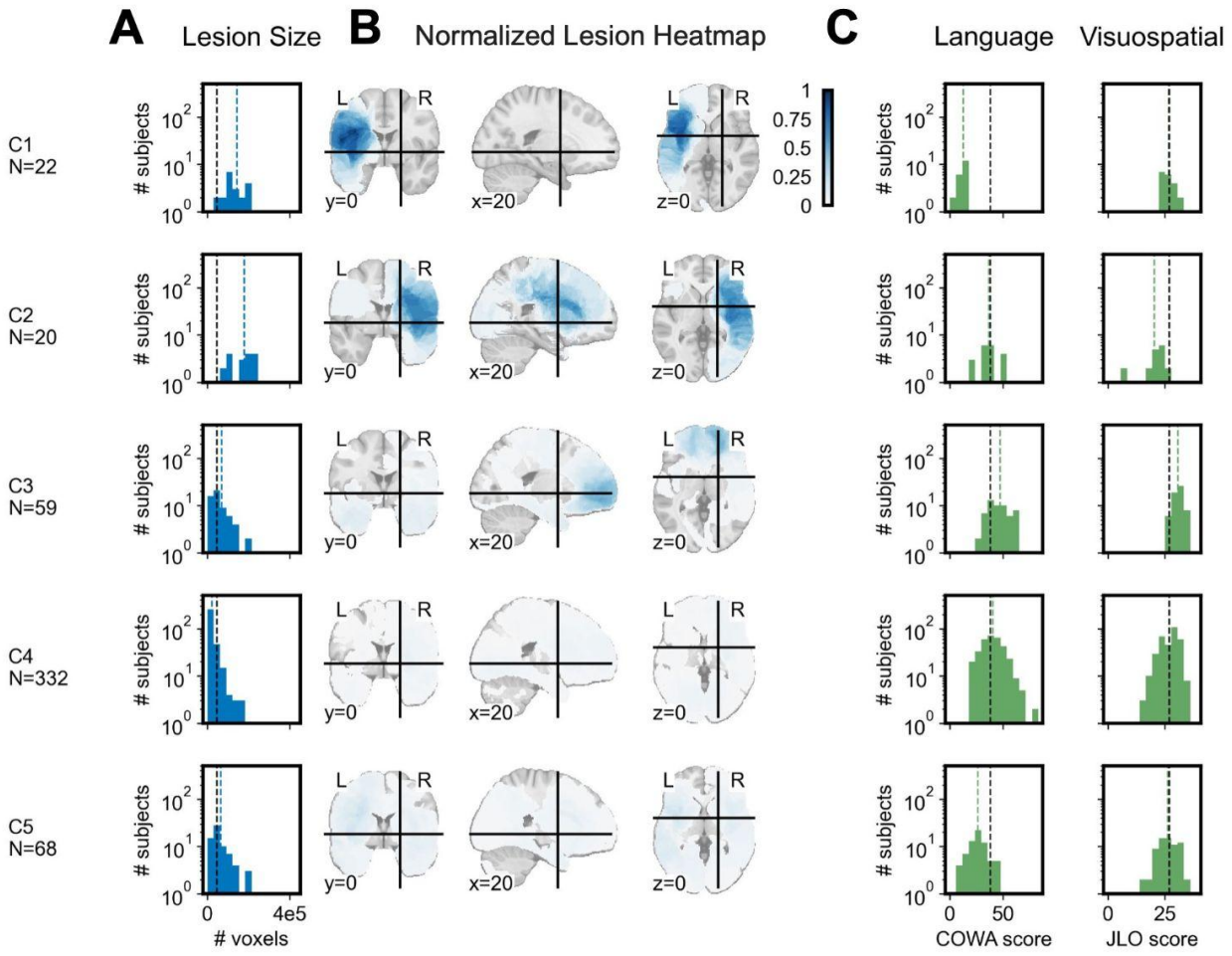


***Figure 4**: CFL finds five lesion categories with distinct performance outcomes on language and visuospatial tasks. A) Lesion size (number of voxels lesioned) distributions associated with individuals in each lesion category, y-axis log scale. B) Normalized lesion heatmap within each category at central coronal, sagittal, and axial slices held constant across categories. Crosshairs indicate the locations of slices in other perspectives. Maximum overlap occurred at the following MNI coordinates: C1 (-42, 0, 15); C2 (31, -19, 30); C3 (17, 46, -10); C3 (31, 2, 7); C5 (-29, 14, 23). Heatmaps across all axial slices can be found in S1. C) Distributions of COWA (left) and JLO (right) scores associated with individuals in each lesion category, y-axis log scale. The cohort mean is shown in black. Within-category mean is shown in green.*

Lesion categories are stable with respect to participant demographics

While CFL preserves causal information present in the micro-level data, it cannot rule out confounding effects without further assumptions or analysis. While the lesion itself is an intervention, the type of lesion that a person has may be correlated with other factors that also affect their behavioral performance, thereby confounding the relation between lesion and behavior. We investigated the potential for confounding of the CFL results with respect to the demographic measures that were available for the data: *age at time of lesion scan, sex, whether-hispanic, race,* and *education*. *Whether-hispanic* and *race* responses were homogenous across the cohort (see S3). To control for the remaining three demographic quantities of *age at time of lesion scan, sex*, and *education*, we compared the discovered categories when CFL is run with vs. without inclusion of these quantities alongside the lesion masks (Figure 5, S3 B, C). If a demographic quantity provided a better explanation than the lesion categories of the measured behavioral outcomes, then its inclusion in CFL would result in a shift of the discovered categories in order to track the demographic quantity, rather than the lesions. The demographic quantities would provide crucial information about the behavioral outcome not already captured by the lesion information – and CFL would detect that.

We found that the lesion categories remained largely unperturbed by the inclusion of demographic information, as shown in the Sankey diagram in Figure 5. We identify the same number of categories partitioning the cohort in essentially the same way. The consistency across the categories suggests that, although these demographic quantities are distributed differently across the five cause categories, the effects do not appear to influence the brain-behavior relationships identified with CFL. The lesion mask already screens off their effect from the language and visuospatial scores (S3 C). Overall, these results are consistent with prior findings that the effects of lesions on cognitive abilities far outweigh any confounding due to demographic quantities[26]. These findings, however, do not preclude the possibility that the relation between lesions and cognitive scores is still confounded by some further *unmeasured* quantity.

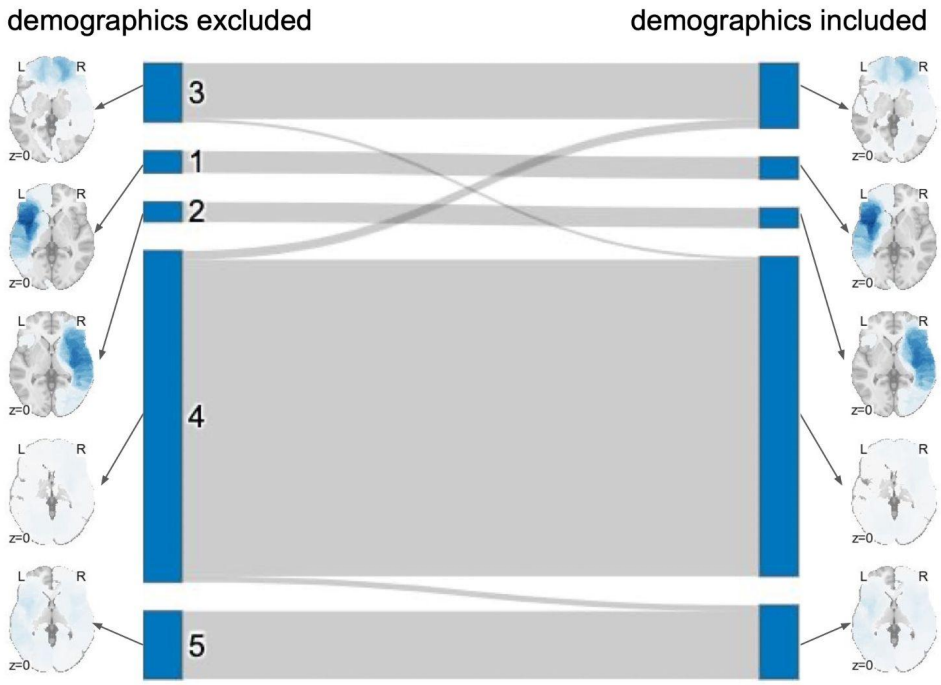


Figure 5*: Comparison of lesion categories found with (right) and without (left) demographic inclusion. The thickness of a given band in the Sankey diagram indicates the proportion of samples that were assigned to category i by CFL when demographics were excluded and category j by CFL when demographics were included. Note that the number of categories could also have changed between these two analyses, but did not.*

Up to this point, conventional behavioral assessment scores were used to evaluate whether CFL could derive expected lesion-behavior relationships (left hemisphere lesions & language impairment, right hemisphere lesions & visuospatial impairment). After demonstrating these expected findings, we now perform the same analysis with depression measures as the outcome. However, unlike the previous analysis, we use responses to individual questions on the BDI questionnaire as the outcome measure instead of an aggregate score. Depression is a construct with distinct processes that is variably assessed using overall severity (e.g. total BDI score) or sub-scores based on distinct questions (e.g. cognitive versus somatic-affective). By applying CFL to individualized response data, we can find the optimal granularity at which to consider both lesions and depressive symptoms for this analysis.

CFL identifies three lesion categories with distinct effects on BDI

Having demonstrated the proof of concept analysis of CFL with respect to largely established conclusions in the case of language and visuospatial ability, we turn next to an area with few established findings: the case of depression. For our second analysis, we looked at the relationship between lesion masks (high-dimensional ‘cause’ variable) and responses to 21 individual questions on the BDI questionnaire (high-dimensional ‘effect’ variable) for subjects in cohort 2. In addition to aggregating lesion data, here we also wanted to investigate the optimal

granularity for the effect measure by coarsening the itemized depression measurements in a principled manner. The CFL algorithm identifies the optimal granularity of the effect space by grouping individual responses to the BDI questionnaire (the effect) together that have similar likelihoods of occurring given the lesion location (the cause). That is, the optimal granularity of a cause-effect relation is identified by coarsening the *cause* when cause categories have similar effects, and coarsening the *effect* when the likelihoods of the effects are similar given the cause (see Methods: Causal Feature Learning (CFL) for details).

The high-dimensional expected conditional probability of BDI responses given lesion location was estimated with ridge regression, and a K-means model was used to partition the space of lesion masks, just as was done in the previous analysis for language and visuospatial ability. The coarsening of the effect was implemented with a second K-means model applied to the estimated probability of each sample in effect space given each cause category found in the previous algorithmic step. All free parameters were tuned in an unsupervised manner and can be found in the Methods section.

CFL identified three distinct lesion categories and three distinct BDI response patterns (Figure 6, see S6 for a detailed discussion of the selection of the number of categories). We note that there need not have been an equal number of cause and effect categories. Cause category C1 identifies lesions localized in vmPFC that have the highest loading on effect category E2, which is associated with a better outcome for the subjects (low to moderate BDI response scores). Furthermore, among the three cause categories, it has the highest loading on E1, the effect category with the best depression outcomes (low BDI scores). In contrast, cause category C2 identifies dlPFC and insula lesions that correspond to the worst BDI outcomes (highest scores), category E3. These lesions, found in about 13% of our cohort, have a strikingly detrimental effect on depression scores. Lastly, cause category C3 contains non-localized lesions corresponding largely to moderate (effect E2) and poor (effect E3) outcomes (moderate and high BDI response scores).

These findings are notably distinct from those we presented in Figure 4, where particular lesion profiles were causally related to *unidimensional* distributions of JLO and COWA scores. While the scores shown in Figure 4 are the commonly used summary scores, here in Figure 6 we are aiming to distinguish unique profiles of item-wise BDI scores.

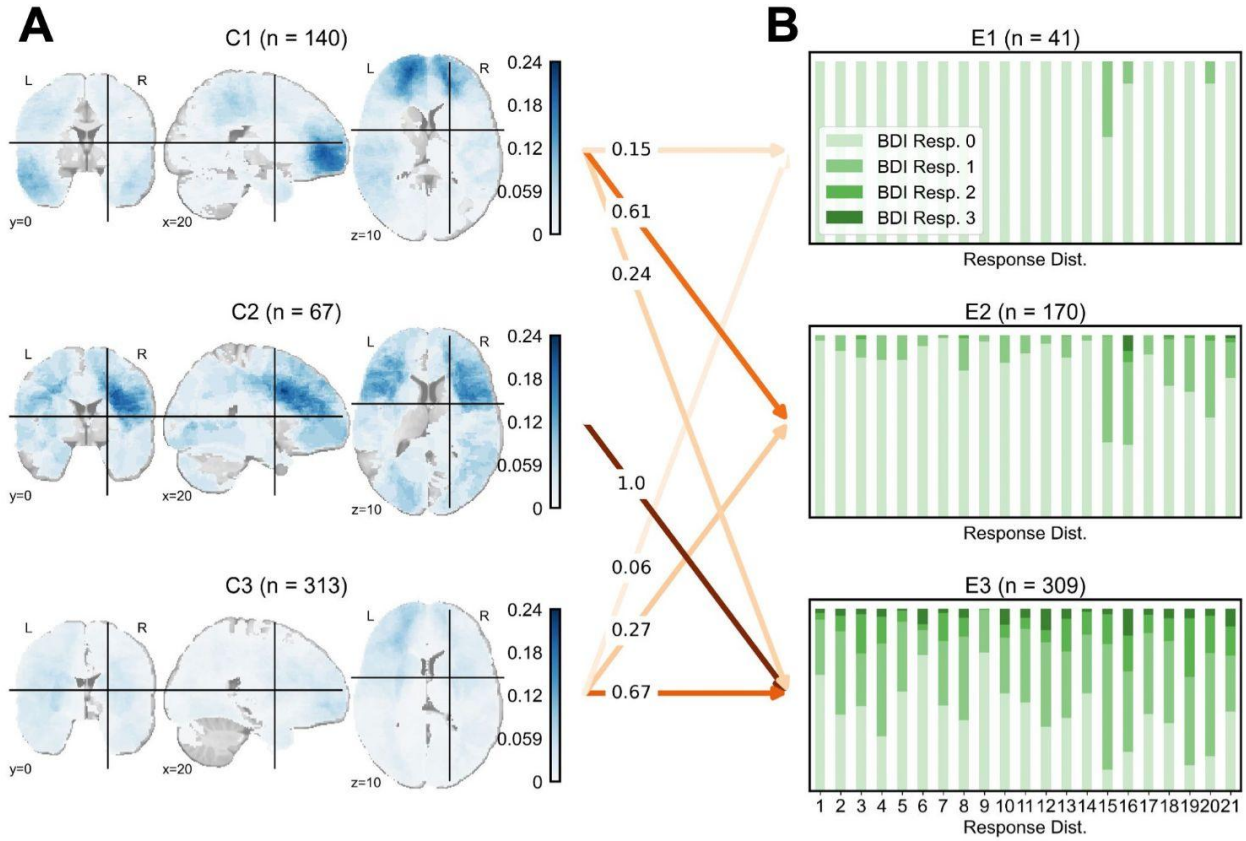


Figure 6: *CFL finds three lesion categories with distinct outcomes over depressive symptoms, and three BDI categories with distinct likelihoods (arrow weights) given any lesion. A) Normalized lesion heatmap within each category at central coronal, sagittal, and axial slices held constant across categories. Crosshairs indicate the locations of slices in other perspectives. Maximum overlap occurred at the following MNI coordinates: C1 (-22, 49, -7); C2 (28, 13, 25); C3 (-30, 26, 19). Heatmaps across all axial slices can be found in S5. B) Distributions of responses to each of the 21 BDI questions within each BDI effect category found by CFL. Orange arrows between panels A and B indicate the probability of each BDI effect category given each lesion cause category.*

The results of this CFL analysis align with the Koenigs et al. 2008[23] findings that recover the same contrast between the two regions with a hypothesis-driven approach. CFL's methodology strengthens these results by going beyond considering a handful of subjects in each of the two groups of interest to instead consider lesions throughout the brain in a much larger sample. In this way, in addition to comparing the relative effects of lesions to the two regions, it also confirms that these are appropriately coarse regions to consider in the first place and thoroughly investigates whether any other regions may be involved in this relationship as well (to the extent that other lesions were represented in the data).

As was done with the language-visuospatial dataset, a demographic analysis showed no evidence of confounding by age, sex, and education. Demographic distributions can be found in Supplementary Figures 7,8.

Interpretation of depression effect categories found by CFL

Among the responses to the 21 questions of the BDI questionnaire, CFL identified in an unsupervised manner three categories of responses that characterize the behavioral responses to the lesions. How should these three categories be interpreted? One natural proposal is to check whether they correspond to three different levels of the overall BDI score (the sum of the questionnaire responses). Alternatively, these categories may be better described in the 2-dimensional space spanned by the cognitive- and somatic-affective-question subscores traditionally used to summarize the 21-question responses. Finally, it is in principle possible that CFL has identified previously unrecognized categories that describe distinct behavioral dimensions that the BDI questions implicitly track. In that case, a data-driven technique to identify these axes is required.

We compare these possibilities in Figure 7. Figure 7A (top left) shows a histogram colored by behavioral response category along the one-dimensional BDI score (here normalized by number of questions for consistency in comparison to other aggregate quantities). It shows a close correlation between the three response categories and the mean BDI score (as colored in Fig 7A, bottom left). When we instead plot the same behavioral categories in the 2D space of the somatic-affective and cognitive scores, we still see organization by behavioral category but little apparent improvement in the separation of these categories (Figure 7A top middle). The CFL effect categories line up along the diagonal that the BDI score also tracks (Figure 7A bottom middle). Finally, we considered a data-driven approach to identifying the dimensionality of the effect space. To construct aggregate quantities informed by the behavioral categories CFL finds, we first computed how important each BDI question was in discriminating between each pair of behavioral effect categories, and then hierarchically clustered questions together with similar discriminatory importance across all category pairs (Figure 7B) (Methods: BDI Question Grouping). The resulting dendrogram suggests that there are three or four groups of questions that drive the distinctions between the three effect categories. The single question 19 about the difficulty in concentrating appears quite distinct from the other questions. We considered groupings of questions both into three and four groups but found no significant difference. For purposes of visualization, we present the three groups marked by the dashed line in Figure 7B. Using the mean of the responses in each of these three groups as the three dimensions of the effect space, we can again plot the effect categories now in the 3D data-derived space (Figure 7A, top right). While there indeed is variation in all three of these dimensions, the coloring according to the three behavioral categories tracks the summary BDI score very closely (compare Figures 7A top right and 7A bottom right).

Quantitatively, decoding the effect category from the 2D and 3D BDI representations does not significantly outperform decoding from the 1D BDI score (Fig 7C; Methods: BDI Aggregate Comparison). Overall then, this analysis provides strong independent evidence that the

one-dimensional total BDI score is in fact an excellent measure of the depression behavior resulting from the lesions. The CFL analysis could easily have resulted in a description of the response behavior that could not have been captured in such a 1D statistic, but in fact, the data strongly supports the application of total BDI as a score for depression behavior. Of course, this assessment is limited by the grain of the measures that were obtained in the first place. Moreover, it is an open question whether much more general or broader questionnaires than the 21 BDI questionnaire or the inclusion of task behavior, would have resulted in a different description of depressive behavior, possibly with different brain areas as causes.

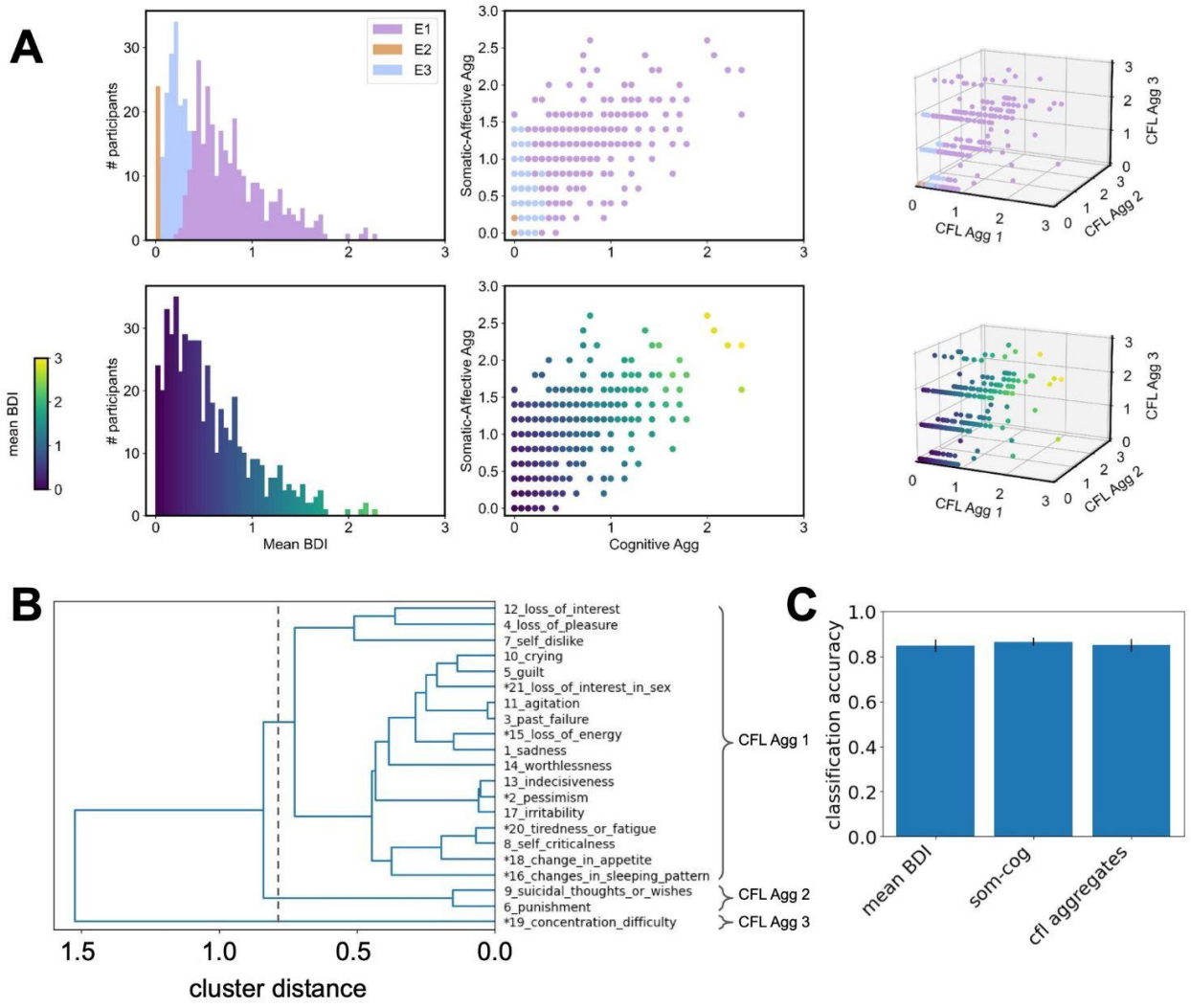


Figure 7*: Comparison of BDI summary statistic candidates. A) Mean BDI aggregate (left), cognitive/somatic-affective aggregates (middle), and CFL aggregates (right), colored by BDI categories found by CFL (top) and by mean BDI (bottom). B) Dendrogram of BDI questions based on hierarchical clustering of how well each question helps in the discrimination of BDI categories. C) Out-of-sample classification accuracy in predicting the BDI category from the three aggregate candidates with logistic regression. The results provide empirical evidence that the one-dimensional summary BDI score is an excellent indicator of depression symptoms as measured by the 21 questions.*

Given that the above analysis suggests that total BDI appropriately summarizes depressive symptoms associated with lesions, we expect that CFL would result in very similar categorization of the lesioned brains when it is run with total BDI as the effect compared to when it is run with the 21 itemized responses to the BDI questionnaire as effect. To this end, we reran CFL with lesion masks as the cause variable and (normalized) total BDI now as the effect variable and compared the resulting lesion categories to those found before when 21-dimensional BDI responses were used as the effect variable (Figure 8). In line with our findings in Figure 7, we found little shift in cause category boundaries between the two effect representations used, confirming that the causally relevant information in the 21-dimensional BDI responses is largely conserved in the total BDI summary statistic.

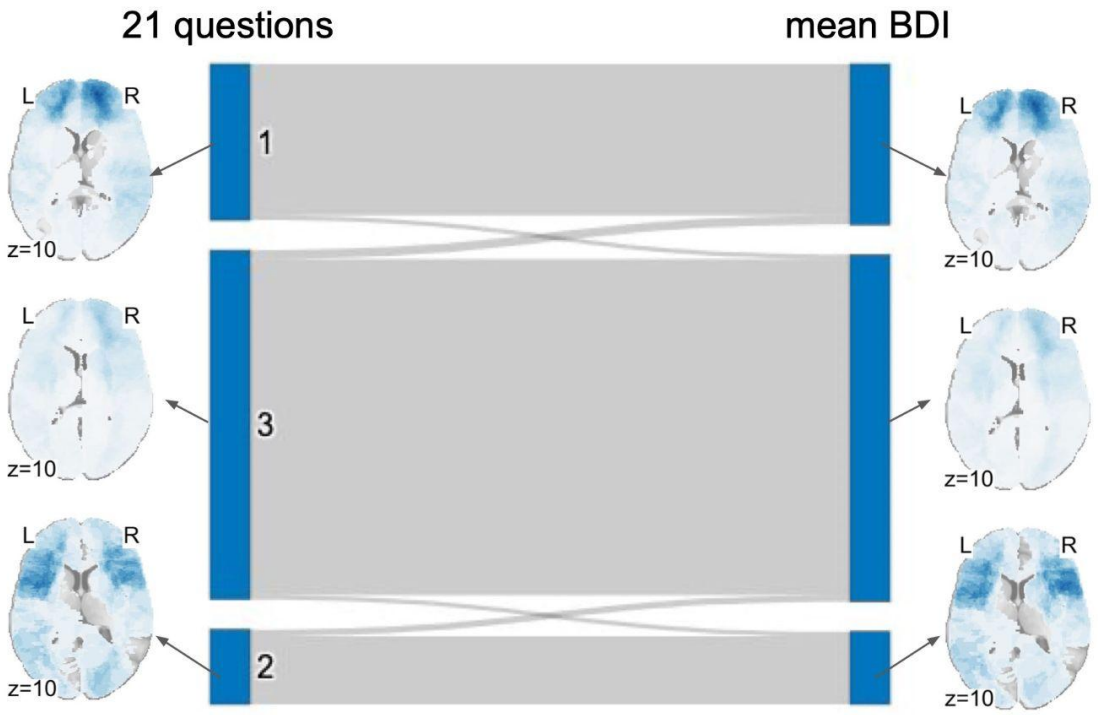


***Figure 8**: Comparison of CFL lesion categories found with an effect variable of 21 BDI questions (left) and of mean BDI (right). The thickness of a given band in the Sankey diagram indicates the proportion of samples that were assigned to category i with the 21-question effect and category j with the total BDI effect.*

Overall, this analysis provides a case study of how to simultaneously aggregate both the lesion map and the behavioral measures to a granularity that optimally describes their relation. The inferred granularity is, of course, subject to the limitations of the availability of varied fine-grained data, both for the lesions and for the 21 questions. However, given these limitations, the analysis empirically confirms the appropriateness of the standard BDI score as a simple sum of the itemized responses. This will not come as a surprise to the practitioner experienced with the use of the BDI score. But to a data analyst, these results suggest that BDI

is a very well-calibrated measure, since CFL could easily have returned a different verdict as its search was in no way restricted a priori to finding aggregates that are just the sum over questions.

CFL lesion categories outperform CCA components in predicting BDI responses

We used the current state-of-the-art, Canonical Correlation Analysis (CCA), as a benchmark for the predictive capacity of CFL. CCA is a method used to identify associations between two potentially high-dimensional variables by learning linear combinations of these variables that maximally align the variation shared between the two[30]. It has been applied to lesion-behavior mapping as a way to identify covariation between lesion features and cognitive test performance. While CCA can relate multi-dimensional cause and effect variables, it does not construct macro-level features explicitly, and it does not capture nonlinear relationships (unless additional components are added). Moreover, CCA is, in essence, a data optimization algorithm, so that the resulting mappings can neglect mechanistic interpretability. We chose to compare CFL to CCA to see how the principled aggregation performed by CFL affects its efficacy as a predictive model.

To quantify CCA as a benchmark, we ran CCA on the same lesion masks and BDI questions that were used in the Cohort 2 analysis and evaluated the model's ability to predict BDI responses from the lower-dimensional lesion representations with 10-fold cross-validation. We assigned a 21-dimensional inner representation to the CCA model, the least restrictive bottleneck possible given that the BDI responses themselves are 21-dimensional.

To compare CFL to this benchmark, we measured the error in linearly decoding the 21-dimensional BDI response from the discrete categories CFL assigned to each lesion mask. We again used 10-fold cross-validation, fitting the CFL model and linear decoder to 9/10ths of the data and measuring error in predicting BDI responses from lesion categories on the remaining tenth of the data.

In predicting BDI responses, we find that CFL's lesion categories achieve lower out-of-sample prediction error (mean MSE = 0.082±0.013) than CCA's 21-dimensional projection of lesion masks (mean MSE = 0.353±0.378) (Figure 10A). Even when we first apply PCA to the lesion masks before running CCA (as recommended in Wang et al.[30]) (mean MSE = 0.133±0.014), CFL still achieves the optimal predictive ability (Figure 10A, "PCA+CCA"). These results are particularly surprising given that CFL achieves this competitive level of predictive accuracy with a 3-state *categorical* representation of the lesion space (the three lesion categories in Fig 6), as compared to CCA which learns a 21-dimensional, continuous bottleneck representation of lesions. This discrepancy demonstrates the advantages of aggregating nonparametrically and aggregating directly on the causal relation itself when predicting behavioral responses from lesion masks. This result offers a reason for caution when interpreting CCA results. Moreover, it suggests that a coarse relation between lesions and depressive symptoms is most optimal for the data available.

When all models are fit using the summary BDI score (rather than the itemized question responses) and their representation of the lesion space is used to predict the BDI score on withheld data, then all three approaches perform similarly (Figure 9B; CCA mean MSE: 0.048±0.030, CCA+PCA mean MSE: 0.036±0.004, CFL mean MSE: 0.042±0.010).

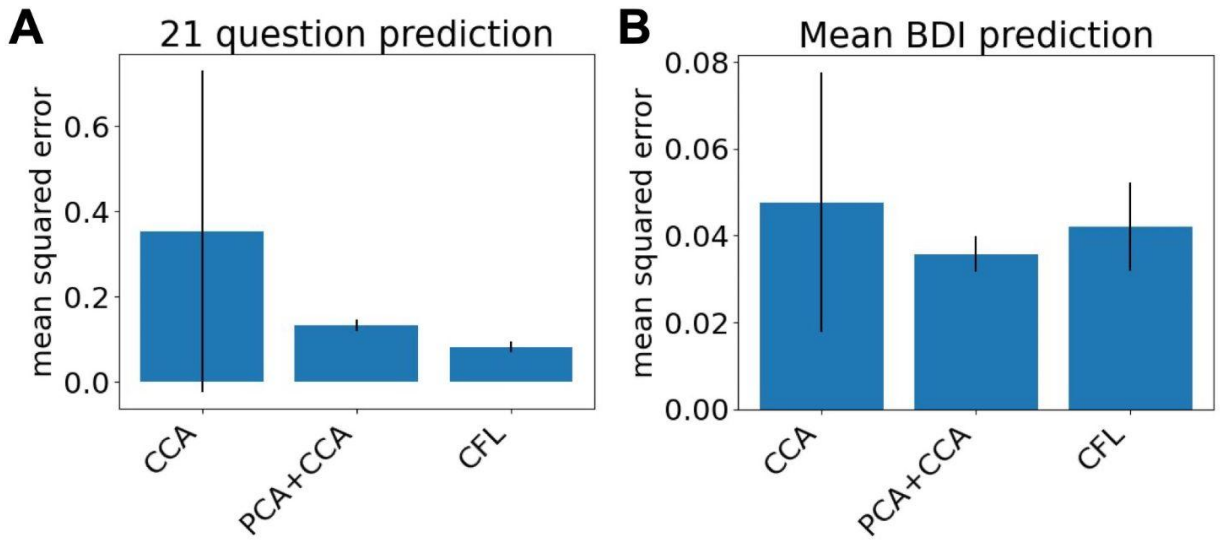


***Figure 9**: CFL achieves better or comparable predictive ability to CCA. Three predictive pipelines are run with 10-fold cross-validation: CCA, CCA with a preliminary PCA step to reduce lesion mask dimensionality, and CFL. A) Prediction error (mean and standard deviation over 10 folds) with the 21 BDI question responses as the effect variable. B) Same as A, except that now the target effect variable is mean BDI. Note that CFL achieves this predictive ability based on three categorical states that lesions are grouped into (Fig 6) as opposed to the 21-dimensional projection of the lesion masks we used for CCA (to match the dimensionality of the output).*

As noted in the introduction, another alternative to CFL is to first coarsen the cause (and possibly effect) space in isolation, and then learn the relation for the resulting clusters. Figure 10 contrasts the lesion macrostates found by CFL (as in Figure 6) with those found by a KMeans clustering model applied directly to the lesion masks (K=5 determined with the Davies Bouldin clustering metric, see Methods: KMeans Clustering Comparison). In contrast to the CFL categories, which group lesions in distinct locations together due to their shared effect on BDI (i.e. category C2 grouping left dlPFC, right dlPFC, and left temporal lesions), KMeans clusters are strictly informed by lesion location and size, leading to the following clustering: smaller (as can be seen from the top right subplot), full right hemispheric, partial right hemispheric, frontal and partial left hemispheric (Figure 10A). While KMeans lesion clusters show some correspondence to BDI, BDI agreement within clusters is comparatively low (Figure 10C), suggesting that the partition drawn by KMeans is grouping some lesions together that do not similarly affect BDI. These differences indicate that even though the number of clusters K used with KMeans was found in an unsupervised manner, coarsening before taking into account the

relation between lesions discarded information that was important to the causal relation by grouping lesions with distinct effects.

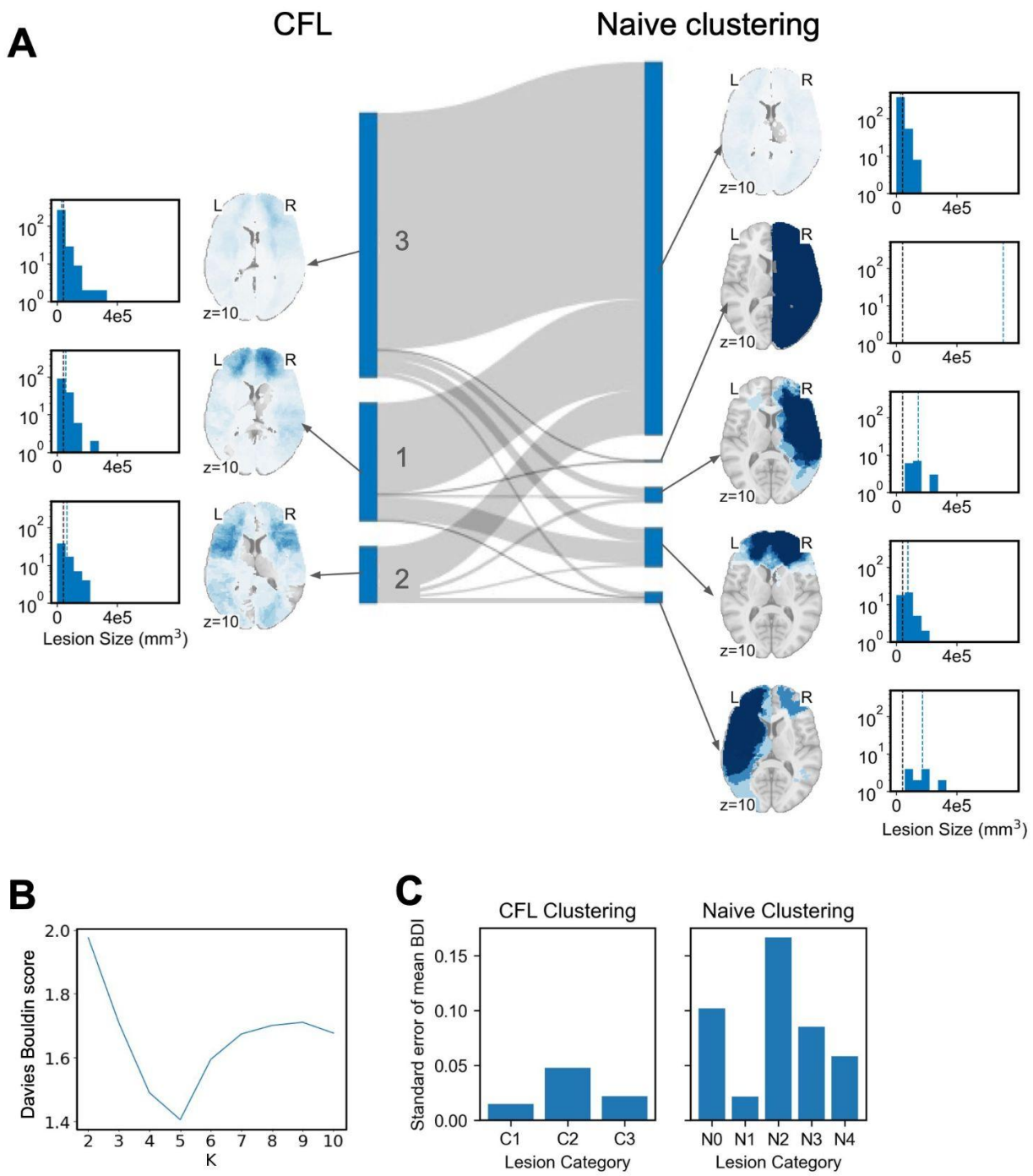


Figure 10*: Comparison of CFL lesion categories to naive clustering of lesions. A) Sankey diagram relating CFL lesion categories to K-means lesion clusters (middle). For each lesion category, the 1) normalized lesion heatmap and 2) lesion size distribution of individuals in the category are shown. B) The Davies-Bouldin goodness-of-clustering metric (lower score is better)*

across varying values of K identifies K=5 as the optimal parameter for K-means during naive clustering. C) Comparison of BDI prediction errors for CFL clustering (left) and naive clustering in lesion space (right). CFL's clustering of the relation preserves important information about the effect of the lesion on behavior that is lost in the naive clustering approach.

Discussion

Summary

We applied a new causal aggregation method, Causal Feature Learning (CFL), to two lesion-behavior mapping questions. First, we used CFL to map lesion location in relation to language and visuospatial task performance and recovered the expected relationships between left hemisphere lesions and language deficits and right hemisphere lesions and visuospatial deficits [26,31–34]. The analysis additionally isolated a relationship between ventromedial prefrontal cortex lesions and unimpaired, or possibly even enhanced cognitive performance relative to the rest of the cohort for these tasks, as also suggested by King et al.[29], identifying an intriguing direction for further research. These results remained stable under checks for potential confounding due to measured demographic quantities.

Having demonstrated that CFL can recover two well-established lesion-behavior relationships, we then turned to the unresolved relationship between lesions and depressive symptoms. CFL found dlPFC and left temporal lesions to disproportionately relate to moderate/severe depressive symptoms, and, surprisingly, vmPFC lesions to primarily relate to non-depressed responses. This categorization of lesions outperformed the low-dimensional description of lesion masks found by Canonical Correlation Analysis (CCA) in predicting BDI responses. It also starkly differed from the lesion clusters found by naive clustering relying on lesion anatomy exclusively, highlighting the importance of taking the causal relation between the lesions and associated behavior into account. Upon comparing the categories of BDI response patterns that CFL identified with the conventional summary statistic for BDI, we found that the total BDI score does in fact summarize the information about the effects of lesions on behavior that is captured by the BDI questionnaire well. This finding suggests that the optimal relationship between lesions and depressive symptoms, as measured in the BDI and supported by the available data, is coarse.

Importantly, the relationships found in both analyses were identified in an unsupervised manner without investigator input. Given a dataset of lesions throughout the brain and the associated full distribution of cognitive test scores, CFL produced these well-established relationships without guidance. The fact that CFL teased apart the dichotomies between left and right hemispheric lesions in relation to language/visuospatial ability, and between vmPFC and dlPFC in relation to BDI [23], lends data-driven support for past work that has recovered the same relationships and is a demonstration of how future lesion-behavior mapping analyses can be accelerated by using CFL. Furthermore, the analysis of the BDI effect space demonstrates how

CFL can be used to relate lesions to fine-grained behavioral data across individual questionnaires and even across broader assessments of behavior that span multiple domains of function.

Limitations

Though lesion studies are much closer to interventional than most human neural data, the observed lesion and measured behavioral performance may still be confounded by unmeasured quantities. We have attempted to control for confounding due to measured demographic quantities, but further unmeasured confounders might have led CFL to form additional distinctions between categories that do not reflect the causal relationship of lesions on behavior. For instance, particular lesion distributions are systematically associated with particular etiologies, such as temporal lobe lesions with epilepsy. Although epilepsy subjects were excluded from this analysis, this caveat still applies to similar sources of unmeasured confounds and must be kept in mind when interpreting results.

Relatedly, given the uncontrolled nature of natural lesion occurrences, the results are biased by the locations of the brain that have been well “sampled” by the lesions. The lesions in our sample of patients did not cover all areas of the brain; lesions varied in size and particular anatomical areas will have been lesioned more often than others. As a result, even though there might be sharply delineated causes of behavior in the brain, CFL may not be able to precisely aggregate to the correct granularity given the available data, and some ambiguity about the optimal granularity can result. Both larger sample sizes and efforts to enrich samples so as to ensure the most isotropic anatomical distribution could help with these limitations. Another note is that certain lesions are systematically excluded – namely, those that result in impairments thought to produce invalid task performances (e.g., blindness would be an exclusionary criterion for performing the JLO), so the effects of those brain regions cannot be appropriately assessed.

A final limitation lies in our ability to visualize and interpret the discovered categories, even once the above limitations have been acknowledged. Though visualizations of average lesion masks and lesion volume distributions allow us to compare lesion location and volume across categories, these methods may not capture other lesion characteristics that may be driving the causal relations that CFL identifies. Further work in visualizing the more subtle lesion characteristics that may constitute the crucial causal distinctions between the identified categories would allow for a more complete understanding of CFL results. A particularly important class of these would be distinctions between damage to gray matter and white matter, where the latter would require further interpretations in terms of disconnection of brain areas as opposed to the absence of particular brain areas.

Future work

As already suggested above, essentially all of the lesions in our study include white matter, and can thus be thought of as producing disconnection syndromes, at least in part. This challenge also points to the broadest recommendation for the future: combine methods. Applying

functional neuroimaging to lesion patients, and combining lesion analysis with white matter atlases and/or diffusion-weighted imaging could all contribute to producing an essential level of description that provides the interface between the lesion masks we obtained and the behavioral impairments with which they are causally associated. That level would help describe the mechanism whereby the lesion causes the change in behavior: through a distributed change in brain function. Indeed, such altered brain function could reflect not only impairments in cognition but also compensation or even improvements [29,35], an intriguing possibility we noted in some of the present findings.

CFL provides an opportunity to systematically evaluate these brain-behavior relationships in a principled way. Just as a variety of different measurements of brain activity could be integrated into CFL on the cause side, a broad characterization of human behaviors spanning cognition, personality, and affect could be used in a CFL model on the effect side. CFL could then be used to search for anatomically informed relationships amongst these dimensions of human behavior that are often investigated independently, providing a new empirical test of existing findings and a data-driven approach to identifying new relations that may not have been hypothesized a priori because they cut across existing accounts of the brain-behavior relations.

Acknowledgments

We would like to thank J. Kahn for their help on the CFL code development. I.W. was supported by a Schmidt Academy Fellowship. [Grant acknowledgement to be completed.]

Author contributions

I.W. conducted the data analysis in close collaboration with F.E.. A.B. and R.A. were involved in the development of the project and contributed to the analysis. A.B. and D.T. collected and curated the data. I.W., F.E., R.A., and A.B. contributed to the manuscript.

Declaration of interests

The authors declare no competing interests.

Resource Availability

Lead contact
Frederick Eberhardt (fde@caltech.edu)

Code availability
Code is available at https://github.com/iwahle/cfl_lbm

Materials availability
This study did not generate new unique reagents.

Methods

Subjects

Participants included 585 individuals from the Patient Registry of the Division of Neuropsychology and Cognitive Neuroscience at the University of Iowa Department of Neurology who met study criteria. The Registry includes patients with focal, acquired brain lesions who completed a range of standardized neuropsychological tests and neuroimaging in the chronic epoch (>3 months) after lesion onset. Inclusion criteria for this study included the presence of a stable brain lesion and completion of specific behavioral assessments, including cognitive testing performed as part of a neuropsychological testing battery and mood assessment using the Beck Depression Inventory-II (BDI-II) [36]. Exclusion criteria included a neurological or psychiatric disorder that preceded the brain lesion, or a history of significant alcohol or substance abuse. The VHIS cohort additionally included 195 subjects drawn from the W.F. Caveness Vietnam Head Injury Study registry, which consists of military veterans who

suffered penetrating head trauma while in combat during the Vietnam War era (1967–1970). Inclusion criteria for this study included individuals with structural neuroimaging of their lesion and having completed a BDI-II in the chronic phase post-injury, defined as >3 months later. All VHIS combat veterans had been declared fit for duty at the time of their enlistment in the military prior to their head injury. In accordance with federal and institutional guidelines, all participants from both cohorts provided written informed consent, and all procedures were approved by the University of Iowa Institutional Review Board and are in accordance with the Declaration of Helsinki.

Cognitive Assessment

Participants in the Iowa Registry were assessed on major domains of cognition, including language and visuospatial ability, using a large battery of different tests (see [36] for additional details of individual tests). For the initial 'proof-of-concept' analysis evaluating CFL, we used the Controlled Oral Word Association (COWA) task of verbal fluency from the Multilingual Aphasia Examination (MAE) (minimum score of 0, maximum score dependent on the number of words the subject can generate in one minute), and the Benton Judgement of Line Orientation (30 questions, score range: 0-30).

Mood Assessment

Mood was assessed using the Beck Depression Inventory-II (BDI-II) [36]. This is a 21-item self-reported questionnaire that evaluates characteristic attitudes and symptoms of depression experienced over the preceding two weeks. The items correspond to affective, cognitive, somatic-affective, and vegetative symptoms of depression that align with the criteria used to diagnose major depression in the fourth edition of the *Diagnostic and Statistical Manual* (DSM-IV). Test-retest reliability is high at 0.93, which suggests robustness against daily mood variations; the validity is further supported by high correlations with Hamilton Depression Rating Scale ($r = 0.71$) and MMPI-D ($r = 0.77$) [37]. If participants had repeat assessments the best performance (highest score) was selected.

Lesion Segmentation

Each participant included in the analysis had a focal brain lesion with visible boundaries evident from research-quality structural imaging from T1 and T2 sequences on MRI. CT was used in rare cases in the Iowa Registry when MRI was contraindicated (n=75 of 585 Iowa subjects across both analyses). Imaging for all subjects was performed in the chronic epoch (>3 months since onset). The anatomical segmentation of lesion borders was traced manually for each subject and brought to a common template space for statistical analyses. The MAP-3 method of lesion tracing involves the manual tracing of lesion borders on a template brain using the lesion depicted in an MRI or CT scan as a guide and has been previously described [38,39]. With improvements in automated methods for transforming brains to a common space, lesions traced after 2006 were manually traced on native T1-weighted scans with FSL [40] and then transformed to the 1 mm MNI152 atlas using nonlinear registration and lesion masking techniques available

in ANTs [41]. Because lesions negatively affect the accuracy of the transformation to MNI space, transformations for unilateral lesions were performed using enantiomorphic normalization [42]. The approach replaces the lesion volume with the voxel intensities from its non-damaged homolog to more closely align the individual's brain with the template brain. Bilateral lesions were transformed by applying a cost function mask to the lesion volume [43], which reduces the influence of voxels within the lesion volume on the transformation process. The spatial transforms were applied to the brain and lesion mask. The anatomical accuracy of the lesion tracing was reviewed in native and MNI space and edited as needed by a neurologist (A.D.B.) blinded to mood ratings.

Data Processing

Lesion masks were taken at 1mm^3 voxel resolution (cohort 1 lesion volume range: 110mm^3-429,508mm^3; cohort 2 lesion volume range: 110mm^3- 772186mm^3; full distributions can be found in S9). To fit a stable model with a relatively low sample size, lesion masks were coarsened from 1mm to 1cm voxel resolution by setting each 1cm^3 voxel value to the proportion of enclosed 1mm^3 voxels that were lesioned. The originally three-dimensional lesion masks were vectorized for all analyses. Only voxels located within the brain volume (as opposed to all voxels in the three-dimensional rectangular prism that encapsulates the brain volume) were included. COWA and JLO scores were independently z-scored, as the raw responses lie in distinct ranges. BDI scores were scaled from the original 0-3 response range to a 0-1 range for all questions, to preserve the relative magnitudes of responses across questions.

Causal Feature Learning (CFL)

We use the CFL framework, as described in Chalupka et al. (2016), to 1) partition the space of neural lesions based on the effects that the lesions have on the fine-grained measures of depressive symptoms, and 2) partition the space of depressive symptoms based on how these symptoms are impacted by any given lesion. The partitions of the lesion and behavior spaces define the lesion and depression categories, respectively.

CFL takes as input the (vectorized) data over the high-dimensional variable X, which takes on values in the domain $\mathcal{X}$ representing the cause space at a micro level, and the high-dimensional variable Y, which takes on values in the domain $\mathcal{Y}$ representing the effect at the micro level. In this particular setup, $\mathcal{X}$ corresponds to the space of all possible lesions as represented by the 1cm^3 lesion mask of the brain, and $\mathcal{Y}$ corresponds to the space of responses to the 21 questions in the BDI questionnaire. The goal of CFL is to coarsen these cause and effect spaces to a granularity that optimally represents the causal relationships we can detect in the micro-level data. To do so, CFL groups together micro-level values that play the same (or very similar) roles in the relationship between X and Y.

First, CFL estimates the conditional probability $P(Y|X)$ from data. However, density estimation is a notoriously challenging task in practice, especially in high-dimensions. We approximate this step by estimating the conditional expectation of the distribution $\mathbb{E}[P(Y|X)]$, as was done in

Chalupka et al. (2016)[44]. This approximation is much more stable in high dimensions, amounting to regressing Y on X with the mean-squared-error measure.

For both the language-visuospatial and BDI analyses, we estimated the expected conditional probability with ridge regression models. Regularization hyperparameter alpha was chosen through 10-fold cross-validation over 11 log-spaced values from 10^{-5} to 10^{5}. The smallest value of alpha for which larger values did not significantly improve out-of-sample performance was chosen. alpha=10 was selected for both the language-visuospatial and BDI analyses. Demographic and mean-BDI follow-up analyses were performed with the same model definition and hyperparameter selection scheme. Models were fit with Python=3.8.15 and scikit-learn=1.1.3.

Partition of lesion space

Given a dataset $(x, y) \in \mathcal{D}$, CFL uses the estimate for $\mathbb{E}[P(Y \mid X = x)] \, \forall \, x \in \mathcal{D}$ (e.g. the expected probability of depression responses given each lesion mask in our dataset) to partition the lesion space $\mathcal{X}$. To preserve the relationship between $\mathcal{X}$ and $\mathcal{Y}$, CFL groups points in $\mathcal{X}$ together that give rise to the same distribution in $\mathcal{Y}$. In practice, we cluster samples $x_1, x_2 \in \mathcal{D}$ together if $\mathbb{E}[P(Y|X = x_1)] \approxeq E[P(Y|X = x_2)]$ (a relaxation of the equivalence relation defined in Chalupka et al., 2016). We used the KMeans method to cluster the conditional probability approximations learned in the previous step.

Selecting number of categories

The KMeans method requires one hyperparameter to be set: the number of clusters (K). To choose K in an unsupervised manner, we first performed this clustering with values of K from 1 to 10 (K=1 requires no clustering - all points are in a single cluster). For each resulting cluster assignment across points, we measured how much information was retained about the conditional probabilities that were clustered (S2A, S6A). We quantified this by taking the least-squares estimate of the conditional probabilities from one-hot-encoded cluster assignments and computing a 10-fold out-of-sample mean-squared error. While prediction error continued to significantly decrease, we increased the granularity of our clustering (increased K), as this was an indication that samples with significantly different conditional probabilities were being grouped together erroneously. Once the improvements in prediction became minimal, we concluded that the bulk of the relational information had been captured by the clustering representation. Models were fit with Python=3.8.15 and scikit-learn=1.1.3.

Partition of behavior space

To coarsen the behavior space $\mathcal{Y}$, CFL groups samples in $\mathcal{Y}$ together if they have the same likelihood of arising given any value in $\mathcal{X}$. In other words, samples $y_1, y_2 \in \mathcal{D}$ are grouped together if $P(Y = y_1|X = x) = P(Y = y_2|X = x) \forall x \in \mathcal{X}$.

To compute the likelihood distribution at a given value of Y in practice, we take advantage of the fact that all $x \in \mathcal{X}$ that are in the same cause macro-state have similar conditional probabilities by construction. Therefore, we only need to approximate the likelihood of each $y \in \mathcal{D}$ given each cause macro-state. For each sample y, we approximate this likelihood with the average distance to its K-nearest neighbors in each cause macro-state[44]. This approximation method relies on the idea that samples from the distribution will be closer to each other in regions where the density is high, relative to where the density is low. Finally, we cluster the samples in this likelihood space to provide a partition of the effect space. Likelihood estimates were again clustered using the K-means clustering method and the aforementioned hyperparameter tuning method to select the value of K.

In summary, the CFL pipeline returns two pieces of information: it assigns each sample in our dataset to 1) a cause macro-state (in this case, a lesion “category”), and 2) an effect macro-state (a depression “category”).

BDI Question Grouping (Figure 7)

Our secondary goal in this analysis was to describe the space of depressive symptoms in an unsupervised manner by grouping BDI questions based on their relationships with lesions. CFL only returns a partition over all possible response patterns in the 21-dimensional space of the questionnaire. In general, such a partition gives little insight into what questions are doing the work to define partition boundaries. To come up with such an organization over the *questions*, we grouped questions together based on how important each question was to define the state space boundaries that CFL constructed.

For each question q_i on the BDI questionnaire, we computed an “importance vector,” where each entry in the vector corresponds to a pair of effect macro-states M_a and M_b, and its value is the KL divergence between the distribution of q_i values in M_a and M_b. Intuitively, this importance vector tells us how distinct q_i responses are between pairs of macro-state - the larger the distance between q_i distributions in M_a and q_i in M_b, the more information q_i tells us about whether an effect sample should be in macro-state M_a or M_b.
We then hierarchically clustered the 21 questions by their importance vectors using Euclidean distance, effectively grouping together questions that play a similar role in discriminating between macro-states. The resulting dendrogram suggested that three or four groups of questions distinctly discriminate between macro-states. No significant differences in decoding the effect category were found between using three or four question groups, so three was selected to simplify the visualization. For each question group, the mean of the responses over each question group was taken to construct three “CFL aggregate” quantities in Figure 7A.

BDI Aggregate Comparison

The three candidate representations of BDI responses (mean BDI, somatic-affective and cognitive aggregates, and CFL aggregates) were quantitatively compared by computing the 10-fold out-of-sample accuracy of logistic regression models fit to predict the effect category

from each candidate representation. Somatic-affective and cognitive question assignments followed those described in Koenigs et al. (2008) [23].

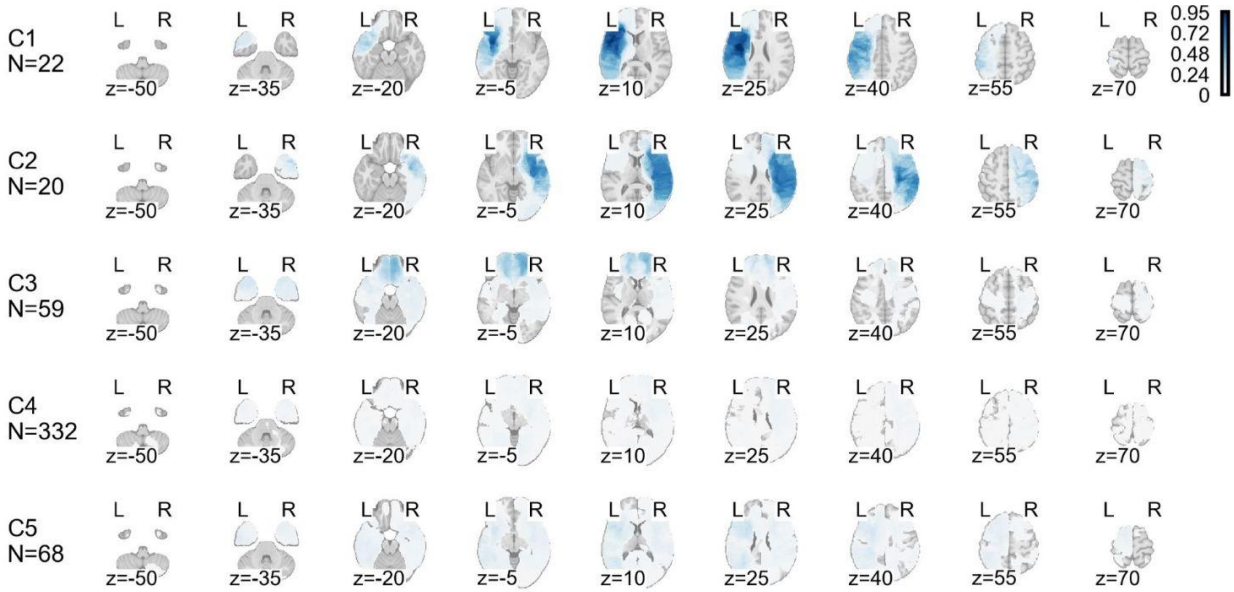
Canonical Correlation Analysis (CCA) Benchmarking

To compare CCA and CFL, we performed two prediction tasks. In the first, we fit a CCA model to map between lesion masks and BDI responses through a 21-dimensional bottleneck. We then evaluated the prediction of BDI values from CCA's low-dimensional representation on a withheld dataset through 20-fold cross-validation.
Per the recommended workflow in Wang et al. 2020 [30], we also modified this CCA evaluation by first using PCA to reduce the dimensionality of the lesion masks down to 21 dimensions before fitting the CCA model.
In the second task, we fit CFL on the lesion masks and BDI responses to learn a partition over the lesion space. We then fit a linear model to predict 21-dimensional BDI responses from the category assignment of the corresponding lesion mask and evaluated the model's out-of-sample performance over 10 folds. The same process was used to evaluate both methods with mean-BDI as the target as well.

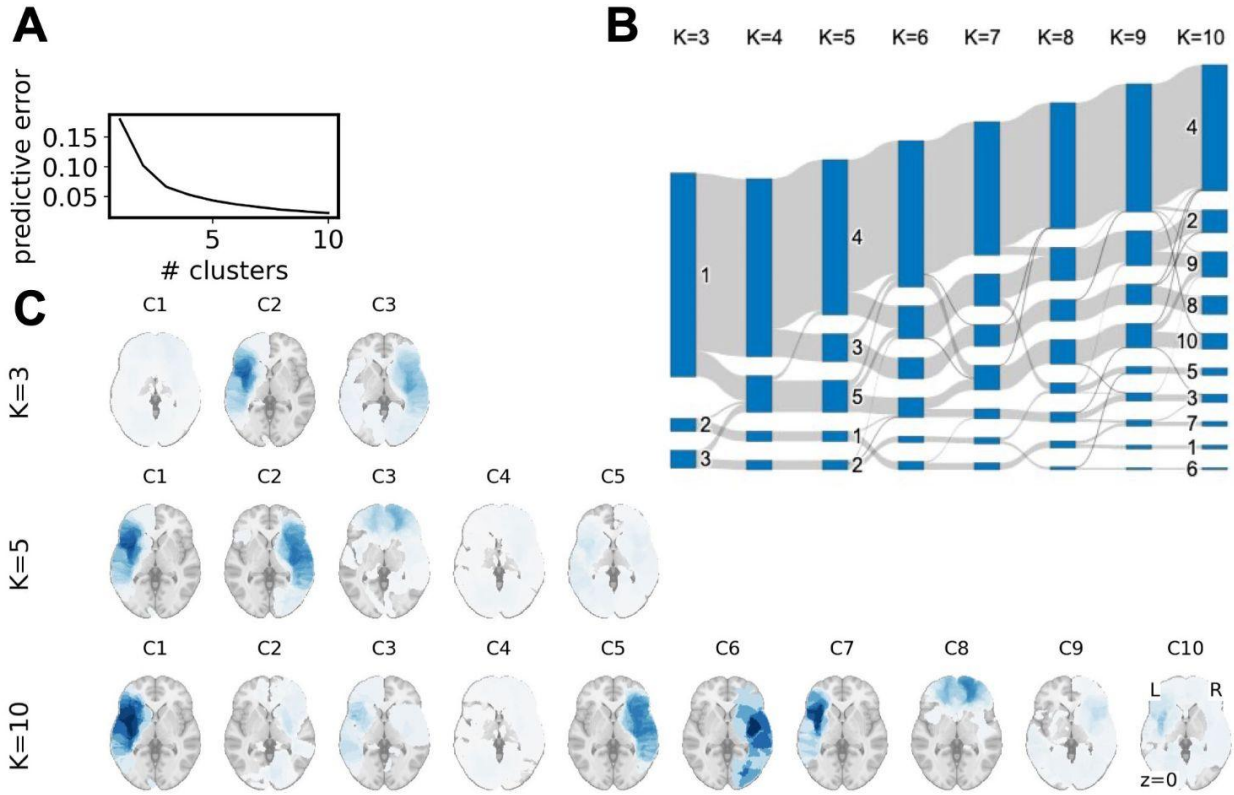
KMeans Clustering Comparison

KMeans clustering was performed on lesion masks over values of K from 1 to 10. The Davies-Bouldin goodness-of-clustering metric was evaluated for each value of K to select the optimal K=5.

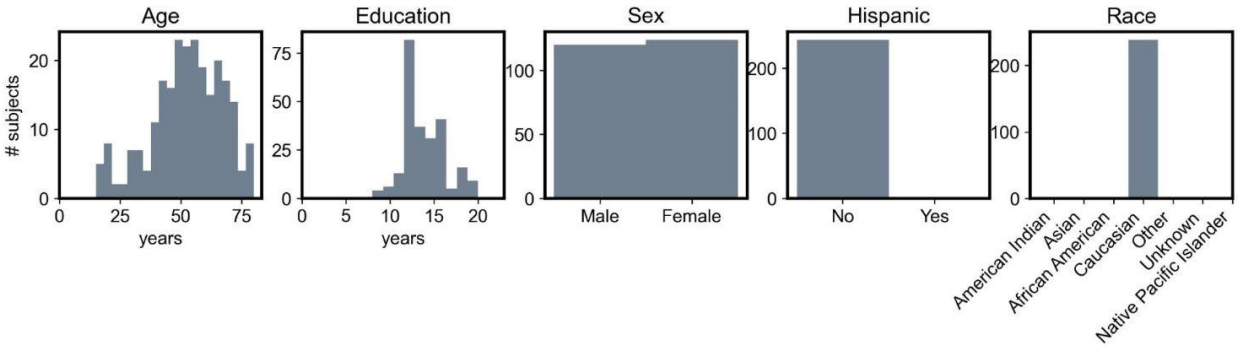
Supplemental information



Supplementary Figure 1*: CFL finds five lesion categories with distinct performance outcomes on language and visuospatial tasks. Axial slices of normalized lesion heatmap for each lesion category relative to language and visuospatial scores.*

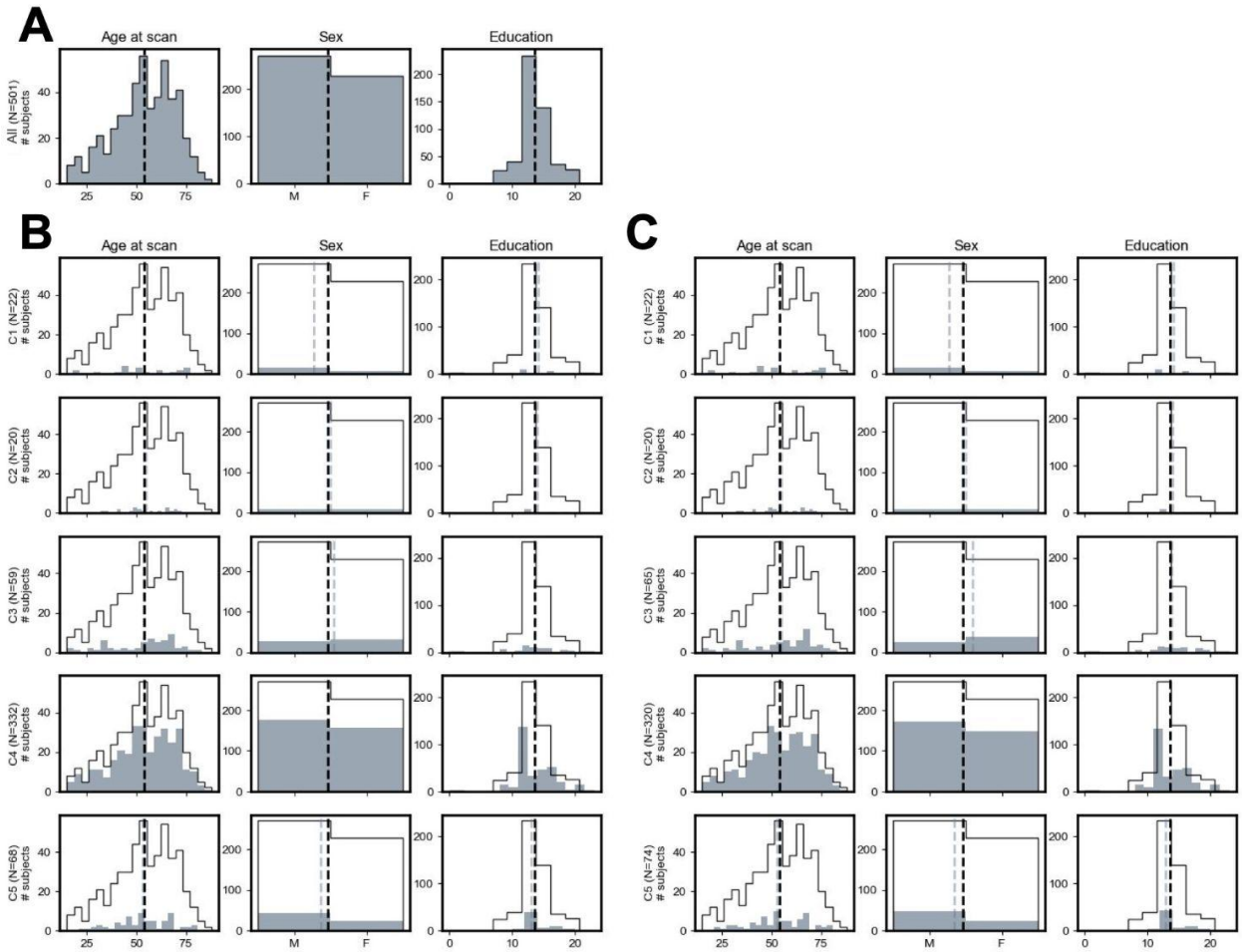


Supplementary Figure 2*: The number of lesion categories* K=5 *is chosen in an unsupervised manner, though there is little variation in results across values of* K*. A) Error in predicting P(Y|X) (the estimated conditional probability of language and visuospatial scores given lesions) from lesion category for different values of* K*. Most error is reduced by* K=5*. B) Lesion category membership across clusterings from K=3:10. C) Normalized lesion heatmap per category for* K=3*, 5, and 10. Identified categories remain qualitatively consistent across values of* K*.*

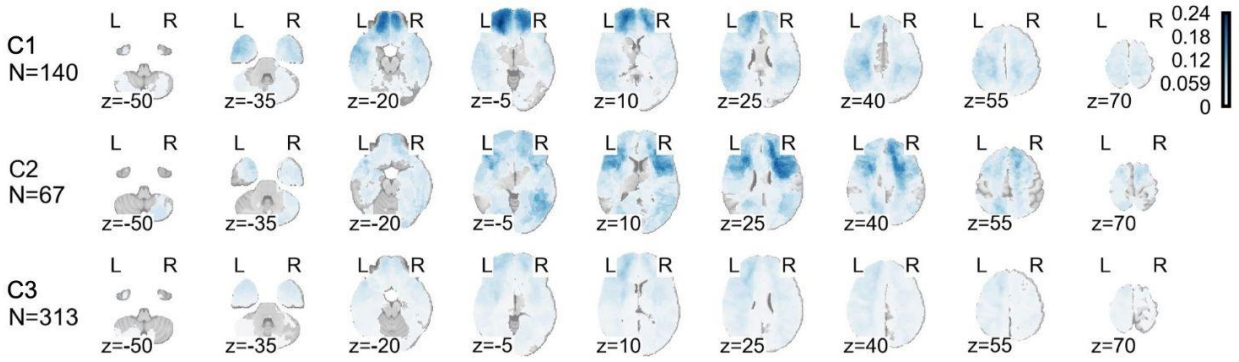


Supplementary Figure 3*: Demographic distributions for 244 out of 501 individuals in Cohort 1 with complete data across the five demographic quantities. No Hispanic or non-Caucasian*

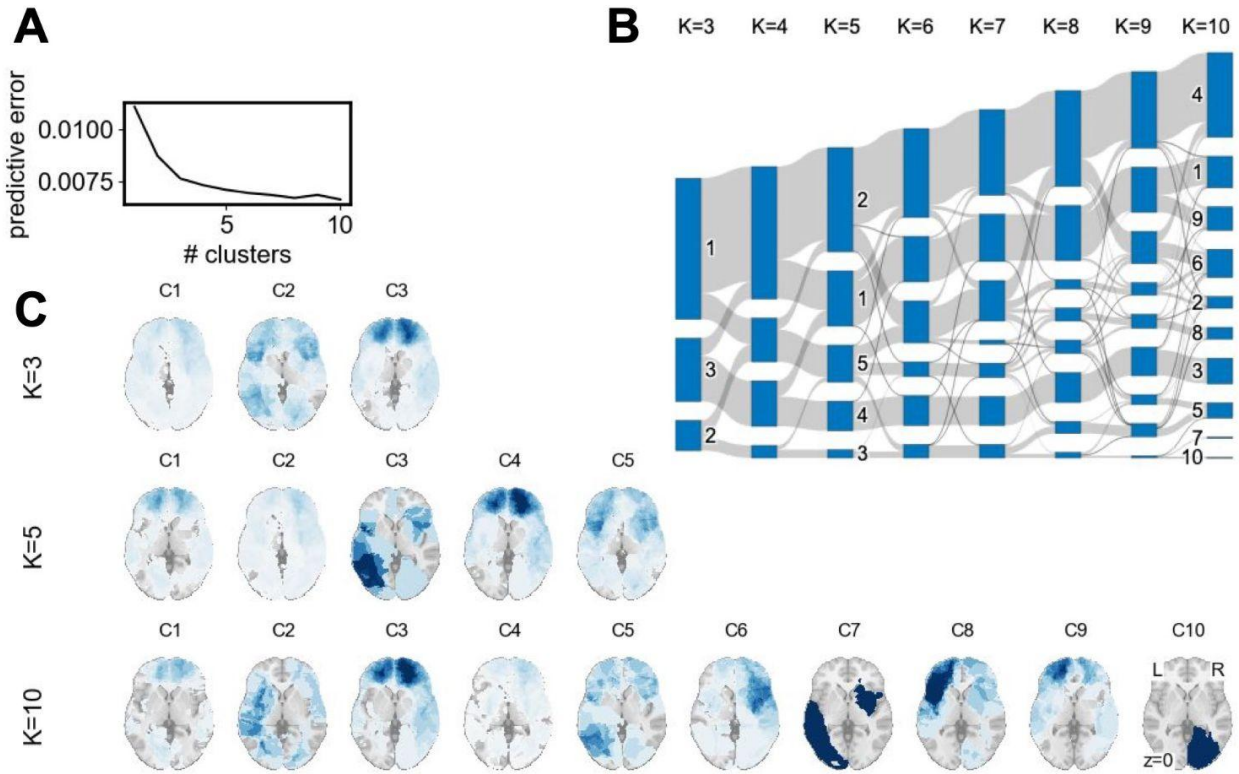
individuals were sampled, so these variables were not considered in the remainder of the analysis.



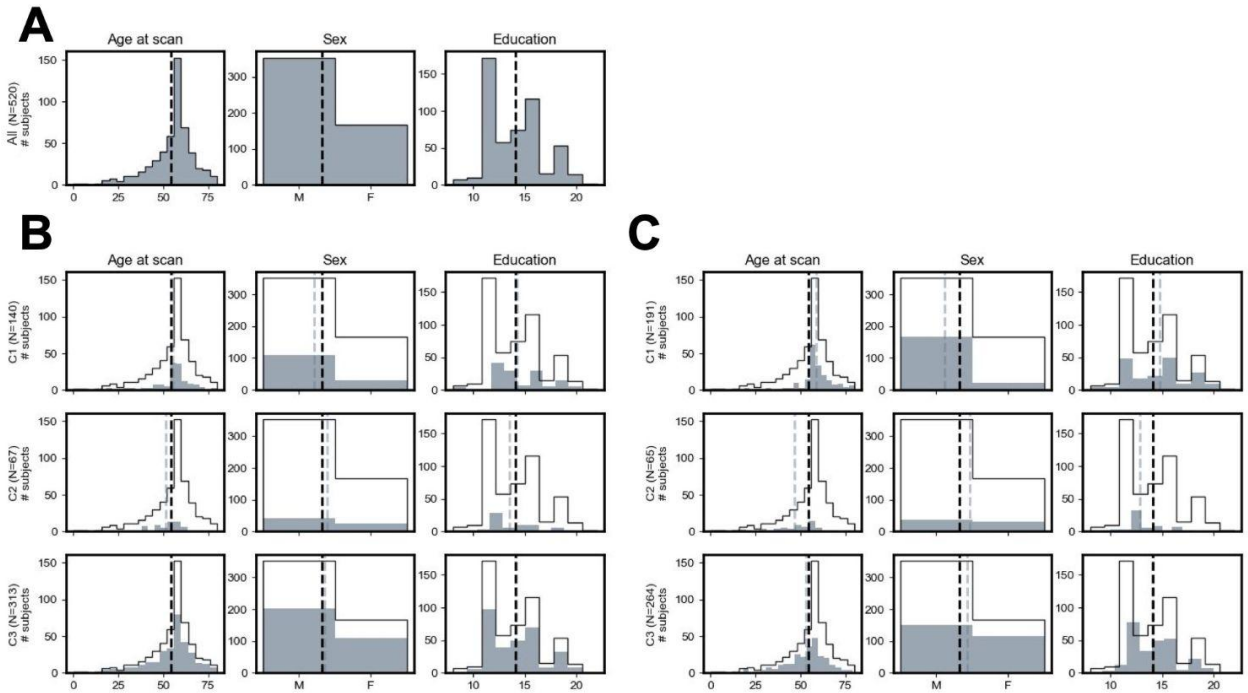
Supplementary Figure 4*: Category-wise comparison of demographics for language-visuospatial analysis. A) Full population demographic distributions. B) demographics per macrostate with demographics excluded from CFL input. C) demographics per macrostate with demographics included in CFL input. Histograms are nearly identical for B and C.*



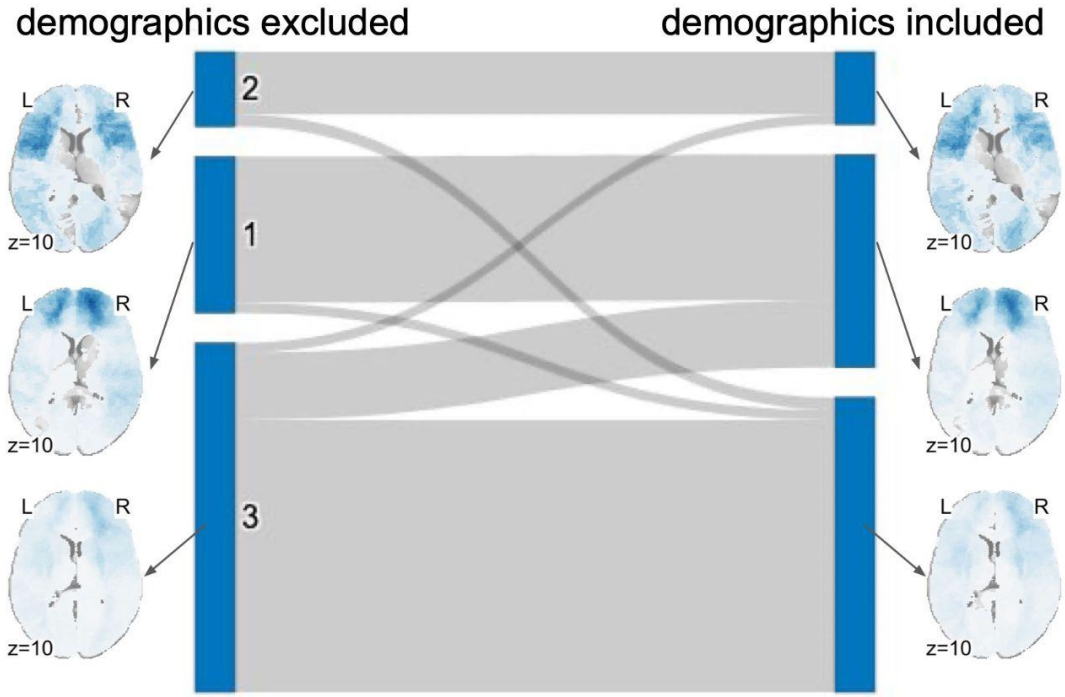
Supplementary Figure 5*: CFL finds three lesion categories with distinct response outcomes across the 21 BDI questions. Axial slices of normalized lesion heatmap for each lesion category relative to BDI scores.*



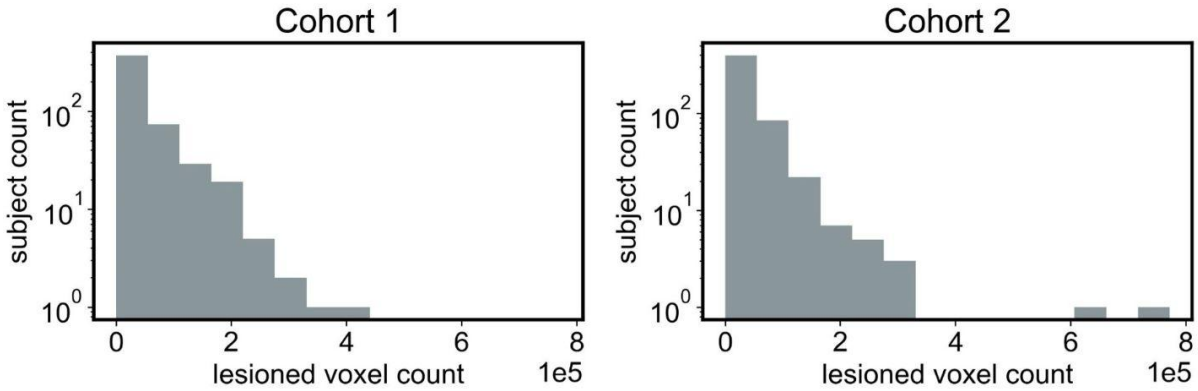
Supplementary Figure 6*: The number of lesion categories* K=3 *is chosen in an unsupervised manner, though there is little variation in results across values of* K*. A) Error in predicting P(Y|X) (the estimated conditional probability of language and visuospatial scores given lesions) from lesion category for different values of* K*. Most error is reduced by* K=3*. B) Lesion category membership across clusterings from K=3:10. C) Normalized lesion heatmap per category for* K=3*, 5, and 10. Identified categories remain qualitatively consistent across values of* K*.*



Supplementary Figure 7*: Category-wise comparison of demographics for the BDI analysis. A) Full population demographic distributions B) demographics per macrostate with demographics* excluded *from CFL input. C) demographics per macrostate with demographics* included *in CFL input. Histograms are nearly identical for B and C.*



Supplementary Figure 8*: Comparison of BDI-analysis lesion categories found with (right) and without (left) demographic inclusion. The thickness of a given band in the Sankey diagram indicates the proportion of samples that were assigned to category i by CFL when demographics were excluded and category j by CFL when demographics were included. Note that the number of categories could also have changed between these two analyses, but did not.*



Supplementary Figure 9*: Lesion size distributions for cohort 1 (left, language-visuospatial analysis) and cohort 2 (right, depression analysis).*

References

1. Broca, M. P. Remarques sur le siége de la faculté du language articulé, suivies d'une observation d'aphémie (perte de la parole). *Bull. Société Anat.* 330–357 (1861).
2. Caramazza, A. On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain Cogn.* **5**, 41–66 (1986).
3. Caramazza, A. & Coltheart, M. *Cognitive Neuropsychology* twenty years on. *Cogn. Neuropsychol.* **23**, 3–12 (2006).
4. Scoville, W. B. & Milner, B. Loss of Recent Memory After Bilateral Hippocampal Lesions. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21 (1957).
5. Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**, 669–672 (1994).
6. Tranel, D., Damasio, H. & Damasio, A. R. Double Dissociation between Overt and Covert Face Recognition. *J. Cogn. Neurosci.* **7**, 425–432 (1995).
7. Goodale, M. A., Milner, A. D., Jakobson, L. S. & Carey, D. P. A neurological dissociation between perceiving objects and grasping them. *Nature* **349**, 154–156 (1991).
8. The Return of Phineas Gage: Clues About the Brain from the Skull of a Famous Patient | Science.
9. Adolphs, R., Damasio, H., Tranel, D., Cooper, G. & Damasio, A. R. A role for

somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *J. Neurosci. Off. J. Soc. Neurosci.* **20**, 2683–2690 (2000).
10. Damasio, A. R. & Tranel, D. Nouns and verbs are retrieved with differently distributed neural systems. *Proc. Natl. Acad. Sci.* **90**, 4957–4960 (1993).
11. Damasio, H., Tranel, D., Grabowski, T., Adolphs, R. & Damasio, A. Neural systems behind word and concept retrieval. *Cognition* **92**, 179–229 (2004).
12. Gläscher, J. *et al.* Lesion Mapping of Cognitive Abilities Linked to Intelligence. *Neuron* **61**, 681–691 (2009).
13. Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L. & Ungerleider, L. G. Discrete Cortical Regions Associated with Knowledge of Color and Knowledge of Action. *Science* **270**, 102–105 (1995).
14. Martin, A., Wiggs, C. L., Ungerleider, L. G. & Haxby, J. V. Neural correlates of category-specific knowledge. *Nature* **379**, 649–652 (1996).
15. Sperber, C., Gallucci, L. & Umarova, R. The low dimensionality of post-stroke cognitive deficits: it's the lesion anatomy! *Brain* **146**, 2443–2452 (2023).
16. Bisogno, A. L. *et al.* A low-dimensional structure of neurological impairment in stroke. *Brain Commun.* **3**, fcab119 (2021).
17. Bates, E. *et al.* Voxel-based lesion–symptom mapping. *Nat. Neurosci.* **6**, 448–450 (2003).
18. Rorden, C., Karnath, H.-O. & Bonilha, L. Improving Lesion-Symptom Mapping. *J. Cogn. Neurosci.* **19**, 1081–1088 (2007).
19. Karnath, H.-O., Sperber, C. & Rorden, C. Mapping human brain lesions and their functional consequences. *NeuroImage* **165**, 180–189 (2018).
20. Chalupka, K., Perona, P. & Eberhardt, F. Visual Causal Feature Learning. *Proc. Thirty-First Conf. Uncertain. Artif. Intell.* (2015).
21. Chalupka, K., Eberhardt, F. & Perona, P. Causal feature learning: an overview. *Behaviormetrika* **44**, 137–164 (2017).
22. Causal Feature Learning, GitHub. (2022).
23. Koenigs, M. *et al.* Distinct regions of prefrontal cortex mediate resistance and vulnerability to depression. *J. Neurosci. Off. J. Soc. Neurosci.* **28**, 12341–12348 (2008).
24. Biesbroek, J. M. *et al.* The anatomy of visuospatial construction revealed by lesion-symptom mapping. *Neuropsychologia* **62**, 68–76 (2014).
25. Tranel, D., Vianna, E., Manzel, K., Damasio, H. & Grabowski, T. Neuroanatomical correlates of the Benton Facial Recognition Test and Judgment of Line Orientation Test. *J. Clin. Exp. Neuropsychol.* **31**, 219–233 (2009).
26. Bowren, M., Jr *et al.* Post-stroke outcomes predicted from multivariate lesion-behaviour and lesion network mapping. *Brain* **145**, 1338–1353 (2022).
27. Nickel, A. & Thomalla, G. Post-Stroke Depression: Impact of Lesion Location and Methodological Limitations-A Topical Review. *Front. Neurol.* **8**, 498 (2017).
28. Beck, A., Steer, R. & Brown, G. *Manual for the Beck Depression Inventory-II*. (Psychological Corporation, 1996).
29. King, M. L., Manzel, K., Bruss, J. & Tranel, D. Neural correlates of improvements in personality and behavior following a neurological event. *Neuropsychologia* **145**, 106579 (2020).

30. Wang, H.-T. *et al.* Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage* **216**, 116745 (2020).
31. Quin-Conroy, J. E., Bayliss, D. M., Daniell, S. G. & Badcock, N. A. Patterns of language and visuospatial functional lateralization and cognitive ability: a systematic review. *Laterality* 1–34 (2023).
32. Benton, A. L., Varney, N. R. & Hamsher, K. Visuospatial judgement: A clinical test. *Arch. Neurol.* **35**, 364–367 (1978).
33. Ng, V. W. K. *et al.* Hemispheric preference in visuospatial processing: A complementary approach with fMRI and lesion studies. *Hum. Brain Mapp.* **10**, 80–86 (2000).
34. Trahan, D. E. Judgment of line orientation in patients with unilateral cerebrovascular lesions. *Assessment* **5**, 227–235 (1998).
35. Kapur, N. Paradoxical functional facilitation in brain-behaviour research: A critical review. *Brain* **119**, 1775–1790 (1996).
36. Tranel, D. The Iowa-Benton School of Neuropsychological Assessment. in *Neuropsychological assessment of neuropsychiatric and neuromedical disorders, 3rd ed* 66–83 (Oxford University Press, 2009).
37. Beck, A., Steer, R. & Brown, G. *Manual for the Beck Depression Inventory-II (BDI-II)*. (P. Corporation, Editor, 1996).
38. Damasio, H. & Frank, R. Three-dimensional in vivo mapping of brain lesions in humans. *Arch. Neurol.* **49**, 137–143 (1992).
39. Fiez, J. A., Damasio, H. & Grabowski, T. J. Lesion segmentation and manual warping to a reference brain: intra- and interobserver reliability. *Hum. Brain Mapp.* **9**, 192–211 (2000).
40. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23 Suppl 1**, S208-219 (2004).
41. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
42. Nachev, P., Coulthard, E., Jäger, H. R., Kennard, C. & Husain, M. Enantiomorphic normalization of focally lesioned brains. *Neuroimage* **39**, 1215–1226 (2008).
43. Brett, M., Leff, A. P., Rorden, C. & Ashburner, J. Spatial normalization of brain images with focal lesions using cost function masking. *NeuroImage* **14**, 486–500 (2001).
44. Chalupka, K., Bischoff, T., Perona, P. & Eberhardt, F. Unsupervised Discovery of El Nino Using Causal Feature Learning on Microlevel Climate Data. *Proc. Thirty-Second Conf. Uncertain. Artif. Intell.* (2016).