

Original Research Articles

Linking public and private neuroimaging data with decentralized analysis via NeuroFLAME Vaults

Bradley Baker, PhD^{1a}, Sandeep Panta, MS¹, Dylan Martin, High School¹, Javier Romero, BFA¹, Ross Kelly, BS¹, Mahmoud Rashidi, PhD², Sunitha Basodi, PhD¹, Ravi Teja Girijala, BS¹, Eimhear Davis, MS¹, Anand Sarwate, PhD³, Jessica Turner, PhD², Frank Hillary, PhD⁴, Kent Hutchinson, PhD⁵, Vince Calhoun, PhD¹

¹ Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Atlanta, GA, USA, ² Department of Psychiatry and Behavioral Health, Ohio State University, Columbus, OH, USA, ³ Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, USA, ⁴ Department of Psychology, Pennsylvania State University, University Park, PA, USA, ⁵ University of Colorado, Boulder, CO, USA

Keywords: federated learning, neuroimaging, data privacy, decentralized analysis, regression, vaults, FAIR data, collaborative neuroscience

<https://doi.org/10.52294/001c.168773>

Aperture Neuro

Vol. 6, Issue SI 1, 2026

Large-scale datasets and global collaborations are increasingly enabling exciting discoveries in neuroimaging research; however, barriers such as data privacy regulations and unequal access to computational infrastructure continue to hinder progress. Federated analysis offers a promising solution by allowing decentralized analysis without centralized data sharing, but logistical challenges often impede its practical implementation. In this work, we present NeuroFLAME Vaults, a novel framework that builds upon the principles of federated analysis to streamline the privacy-preserving, decentralized analysis of neuroimaging data. NeuroFLAME Vaults provide continuously accessible, preprocessed datasets hosted at local sites, enabling researchers to participate in collaborative studies without transferring sensitive data or coordinating synchronous computation. We describe the design and implementation of the platform, demonstrating its utility through multiple vaults derived from public and private datasets, including structural and functional MRI data from studies on schizophrenia, substance use, and neurodevelopment. Vault-based analyses leverage standardized pipelines and ridge regression models to enable cross-site modeling and data harmonization. Our results replicate known findings in neurodevelopment, psychiatric disorders, and substance use, validating the approach and showcasing its potential for inclusive and scalable neuroimaging research. NeuroFLAME Vaults significantly reduce the technical and logistical burdens of collaborative analysis, while upholding FAIR data principles, offering a powerful tool to accelerate global neuroscience.

1. INTRODUCTION

Over the past decade, neuroimaging data collection has accelerated significantly through large-scale, global initiatives such as the Adolescent Brain Cognitive Development (ABCD) study (>10,000 participants),¹ the UK Biobank imaging extension (targeting 100,000 brain scans),² and the Human Connectome Project (HCP).³ Additionally, platforms like OpenNeuro⁴ have provided access to a wide array of datasets, now hosting over 600 public neuroimaging datasets representing more than 20,000 participants across multiple imaging modalities.⁵ These resources offer transformative potential for neuroscience and mental health re-

search. However, two persistent barriers continue to hinder large-scale collaborative data analysis: (1) strict privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation), and (2) unequal access to computational infrastructure, especially among under-resourced or internationally distributed institutions.⁵ Federated analysis platforms share intermediary statistics to compute analyses rather than full data sets, offering one solution to these barriers; however, further issues arise due to logistical barriers involved in coordinating simultaneous federated runs. It is our goal in this work to illustrate how the vaults feature in the new NeuroFLAME (Neuroimaging Federated Learning & Analysis for Multi-Site Environments) software en-

^a **Corresponding Author: **

Bradley Baker, PhD (bbaker43@gsu.edu)

Tri-Institutional Center for Translational Research in Neuroimaging and Data Science
Atlanta, GA, USA

environment addresses data sharing issues in neuroimaging analyses while simultaneously providing federated analysis resources that circumvent the frequent logistical barriers involved in standard federated analysis frameworks.

In this work, we distinguish between federated learning, federated analysis, and vault-based collaboration. Federated learning traditionally refers to distributed optimization procedures in which model updates are exchanged between actively participating sites. Federated analysis more broadly includes decentralized statistical workflows that exchange summary statistics or model parameters without centralizing raw data. NeuroFLAME Vaults extend these paradigms by introducing persistent federated nodes that remain continuously accessible to approved collaborators, thereby reducing the coordination burden typically associated with synchronized federated studies.

This article outlines how our new decentralized data sharing and analysis platform, NeuroFLAME, and the *vaults* feature address these challenges and more in a fashion specifically tailored for the decentralized analysis of neuroimaging data. We will present in detail how data analysis with vaults works, including the standardized pre-processing pipelines that were run on the data. The results section shows some of the analysis results from the data sets in the NeuroFLAME vaults. The following section presents the motivation behind the development of the vaults feature and its utility for global collaborative Neuroscience research.

1.1. THE CASE FOR FEDERATED PLATFORMS IN NEUROIMAGING

Federated learning is a specific type of decentralized data analysis framework that focuses on collaboratively training machine learning models across devices while keeping data local. A similar but distinct kind of framework for federated analysis involves the sharing of summary statistics or model parameters after model training has been performed locally. Both approaches are rapidly gaining momentum in neuroscience and healthcare as a critical method for enabling privacy-preserving, collaborative analysis across multiple institutions. In particular, federated solutions offer a solution to collaborative analysis of medical data in which data samples are not transferred between sites. Often accompanied by privacy guarantors like secure multi-party computation or differential privacy,⁶ federated algorithms provide researchers with the ability to integrate data sets across studies without the transfer of full samples between collection sites, providing the benefits of increased sample size and diversity while endeavoring to remain compliant with data privacy regulations.

According to a comprehensive systematic review of federated analysis applications in healthcare,⁵ as of August 31st 2023, only 32 out of 22,693 articles contained studies with real-world clinical applications, a mere 0.14%. The vast majority of the reviewed papers were instead proof-of-concept studies without large-scale real data applications. Furthermore, only 11 studies featured international collaboration, with a few others limited to local efforts or undisclosed sites. This distribution further reflects the limited

real-world uptake of federation in healthcare and the need for more robust, collaborative platforms. There is clearly a significant gap between theoretical development and practical implementation of federated learning and analysis platforms.

One persistent barrier to federated collaboration in neuroimaging is the significant logistical overhead involved in coordinating real-time federated workflows. In this work, we introduce **NeuroFLAME vaults**, a new kind of federated node that provides persistent access to neuroimaging datasets without the need for active monitoring or coordination by the site hosting the data. NeuroFLAME vaults allow sites to share data in such a way that only the intermediate statistics, such as parameter gradients or final parameters, needed by a given federated learning or analysis algorithm ever leave the site. The NeuroFLAME software, built on the NVIDIA FLARE backend,⁷ supports the vault construct by automating the sharing of these statistics when accessed by other federated clients. Building on previous work introducing federated vaults in the COIN-STAC platform,⁸ we present several new vaults that expand the selection of available analyses and data.

1.2. WHAT ARE VAULTS, AND WHY DO WE NEED THEM?

NeuroFLAME vaults are *persistent, automatically-managed*, federated nodes which host shared data-sets either fully available for public access, or shared with a subset of trusted collaborators, with the sharing controlled entirely via standard federated analysis and learning paradigms. For example, suppose a federated learning algorithm for training a regression model uses distributed stochastic gradient descent where each update to local models involves sending parameter gradients belonging to a particular batch or group of batches. Where a standard federated study may require explicit communication between participating sites to make sure all clients are active, a federated computation using a vault will be able to access the needed gradients without the need to explicitly coordinate with the host of the vault beforehand. While vault access may be restricted to particular users by the owner, once access is provided, the data becomes available for use in approved algorithms for federated analysis or learning, facilitating federated “sharing” without the need for direct coordination.

Vaults represent a new approach to building federated consortia in which participants in federated consortia do not need to actively manage the inclusion of their data in new analyses, offering their data for use by other collaborators entirely automatically through the NeuroFLAME software. Importantly, from the perspective of a given federated algorithm, vaults are treated no differently than normal participants in the consortium, so existing infrastructure, codebases, and regulations can work seamlessly with any site hosting a vault. For external users not hosting a vault, the NeuroFLAME software offers users the ability to interact with these persistent vaults without requiring direct communication and coordination with the vault host, thus reducing the need for additional communication over-

NeuroFLAME Vault vs. Normal Consortium

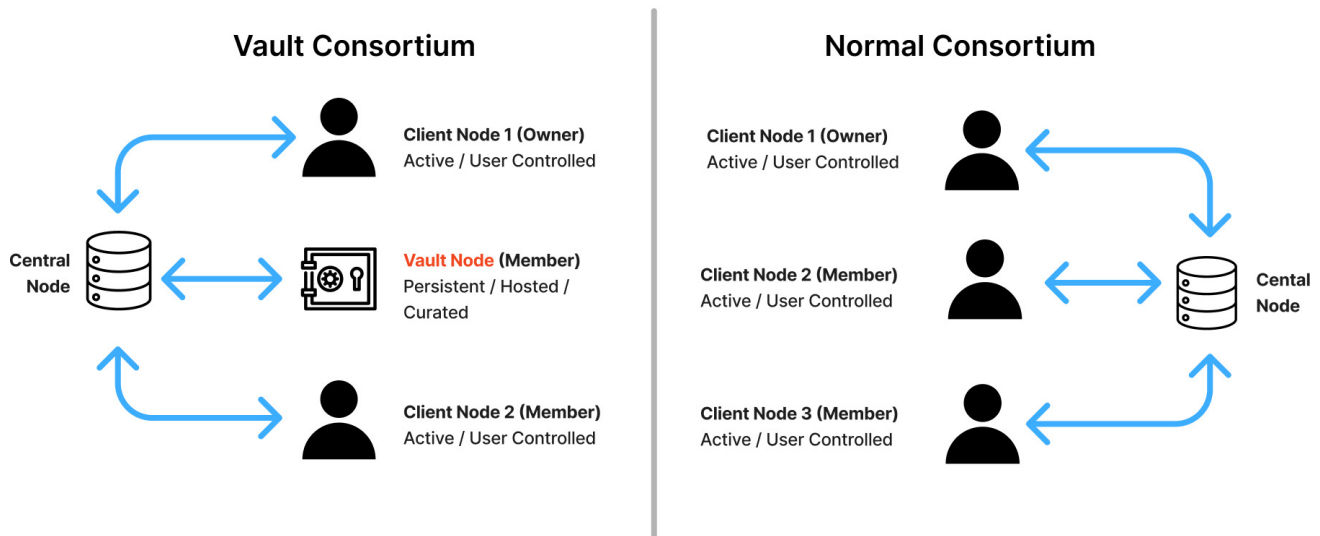


Figure 1. A comparison of vault-based federated consortia with standard federated consortia. When a host provides a vault in the NeuroFLAME software, the data hosted within that vault becomes usable by approved consortium members for use in computations as if the vault host were participating as a normal consortium member. In contrast to standard federated analyses, however, the vault owners have provided a persistent service that does not require additional management or coordination from their team, thus significantly lowering the barrier to access for approved groups that hope to access these data.

head, a persistent barrier to existing federated collaborations.⁸ Figure 1 provides a comparison of vault-based federated consortia with a standard federated consortium. The principal difference is that the vault owner does not need to actively coordinate the integration of their hosted data into the federated computation; the NeuroFLAME software provides access to this vault data as a service, allowing access to approved users without additional overhead from data managers at the vault host's site.

Lack of access to curated and well-annotated neuroimaging datasets remains a significant barrier to collaborative research due to a range of practical, cultural, and infrastructural challenges. Many researchers lack experience with data organization and sharing workflows, and the effort required to prepare datasets for reuse — including documentation, metadata standardization, and ethical/IRB compliance — disincentivizes broad distribution and limits access for secondary analyses. Additional obstacles include the lack of credit structures for data sharing, resource demands for data curation, and heterogeneity across repositories in formats and access procedures, which complicate dataset discovery and reuse.⁹ These barriers reduce the accessibility of curated neuroimaging resources, impede reproducibility, and motivate models like federated analysis or data vaults that enable collaborative research without requiring full data transfer. Vaults help address these barriers by automating the work of persistent data management for collaborative learning. In particular, within the NeuroFLAME software, the hosted vault acts as a service that waits for connections from approved federated collaborators and then provides the required intermediate deriv-

atives needed to perform a given federated analysis. For example, in the regression vaults discussed later in this work, a user accessing a vault via the NeuroFLAME software would be provided with a secure connection to the vault host as if the host was a normal federated node, and the estimated parameters computed on the vault data would be shared from the host to the consortium participant, allowing that participant to perform a federated analysis using the vault data, without communication of full data samples between sites, and indeed without requiring additional coordination from human coordinators at the vault site.

Privacy via Local Storage. NeuroFLAME vaults are designed to support privacy-preserving workflows by enabling persistent decentralized analyses without requiring raw data to be shared or pooled. Data remain stored locally within the vault host institution or hosted on a cloud service when desired, and only computation results or intermediate statistics — such as local regression parameters, summary statistics or model updates — are transmitted when the vault is accessed, according to the needs of the underlying federated computation. This approach to persistent data sharing is grounded in federated analysis principles and allows collaborative research while minimizing privacy risks through the application of quantitative privacy guarantors such as differential privacy.⁶ As vaults rely on the underlying NeuroFLAME software for federated computation, communication between vaults and the central system is encrypted, and privacy-preserving techniques like differential privacy and secure multiparty computation (SMPC) are supported, as with other federated computations. The underlying NVFlare infrastructure uses secure

communication channels, such as gRPC and TCP with SSL certificates, to establish identities and secure communication between participants. A provisioning system creates security credentials for the secure deployment of FL applications in real-world settings. Computations using NeuroFLAME vaults also maintain detailed audit logs of all computations, ensuring transparency, accountability, and compliance with ethical standards. Importantly, vaults do not depend on a specific, quantifiable notion of privacy or a particular threat model, and should in principle be usable with any additional privacy constraints imposed by a particular federated learning or analysis algorithm.

Synchronized standard Pipelines. NeuroFLAME vaults provide standardized analysis pipelines that are accessible through an intuitive graphical user interface (GUI, see [figure 2](#)), cataloguing the available vaults. Data in these vaults is preprocessed using pipelines vetted by neuroscience experts and is made analysis-ready for the users. Unlike traditional methods requiring downloading data from various sources and running local analysis, vaults offer a streamlined, reproducible, and user-friendly alternative. This fosters collaboration across researchers of varying technical expertise.

Vault datasets: FAIR research. NeuroFLAME Vaults' UI displays all relevant information necessary for performing analyses using the respective vault data. This information includes, but is not limited to, dataset descriptions; scanning and collection protocols; data modalities; demographic and other variables; details of preprocessing and quality control pipelines; URLs of research publications related to the dataset; and the types of analyses available to run on the vault data. This information helps researchers choose the relevant vault according to their project needs. Vault analyses run on users' local machines inside Docker and Singularity containers, which provide isolated, consistent environments for running software. These containers package the analysis code along with all required dependencies, ensuring that analyses can be executed reliably across different systems without conflicts. This containerized approach makes analysis results reproducible, as the Docker and Singularity containers maintain the same software environment regardless of where they are run. Once users are ready to publish analysis results utilizing a Vault dataset, they will be provided with standardized references pertaining to NeuroFLAME Vaults. Reproducing a previously published analysis on a vault is made simple by following tutorials provided in the NeuroFLAME documentation (n.d.) and through the use of a wizard built into the software itself.

Combining Local and Vault Data for Diverse Research. Current neuroimaging data sharing platforms generally do not allow users to upload their own data as persistent, "always-accessible" vaults. This means that datasets that remain available for federated analyses only as long as the computer hosting the data is online, becoming inaccessible when the user logs off. This limitation restricts participation by institutions with sensitive, small-scale, or hard-to-access datasets. As a result, racially, ethnically, geographically, and socioeconomically diverse groups could

potentially remain underrepresented in neuroscience research, limiting the generalizability of research findings. NeuroFLAME's vaults address this gap by enabling these institutions to securely host their data as vaults that are continuously accessible for collaborative research. By facilitating seamless integration of local datasets with vault-hosted data, NeuroFLAME vaults enhance inclusivity, especially in studies focusing on populations traditionally underrepresented in neuroimaging research. Additionally, large, curated datasets in vaults facilitate cross-study validation, helping identify biases and anomalies. This strengthens the reliability and robustness of research outcomes, and follows the guiding principle of previous collaborative efforts like the ENIGMA consortium.¹⁰

Public and Private Vaults in NeuroFLAME. Researchers can store both public and private data in vaults on their machines using our standardized vaulting process. Vault owners retain full control over who can access and run computations on their data, and only approved pipelines are allowed.⁸ Furthermore, hosting of vaults is inherently flexible. Whether the data meant to be shared is hosted on premises or on a cloud solution like an AWS EC2 instance, all that is required to host a vault is an internet connection and the NeuroFLAME software. Although some initial coordination may be required to ensure that the chosen hardware and software are approved by the site hosting the vault, the NeuroFLAME software handles subsequent communication automatically.

2. RELATED WORK

In the last decade, a range of data sharing models have emerged within the neuroimaging ecosystem, allowing researchers to access large datasets and perform local analysis. Some of these platforms allow users to run collaborative analysis. In the following section, we outline some of the major relevant platforms and address their ability to run decentralized privacy-preserving data analysis.

The ENIGMA consortium is perhaps one of the earliest efforts for widespread, privacy-sensitive collaborative data-sharing for neuroimaging.¹⁰ Studies led by the ENIGMA consortium pioneered the sharing of summary statistics between distinct data-collection sites for the purpose of furthering large-scale, collaborative meta-analyses utilizing various modalities of neuroimaging and omics data. Many of the principles in contemporary federated neuroimaging aim for the same goals of large-scale, privacy-sensitive collaboration as these early efforts; however, federated analysis and learning provide a more general framework for collaboration, which is not inherently tied to meta-analysis. In fact, federated platforms like COINSTAC and NeuroFLAME can integrate seamlessly with the meta-analyses pioneered by the ENIGMA group.¹¹

OpenNeuro is a free, open platform for sharing and accessing neuroimaging data, designed to support transparency, reproducibility, and collaboration in neuroscience research. It primarily hosts datasets organized using the BIDS (Brain Imaging Data Structure) standard,¹² covering various modalities including fMRI, EEG, MEG, and struc-

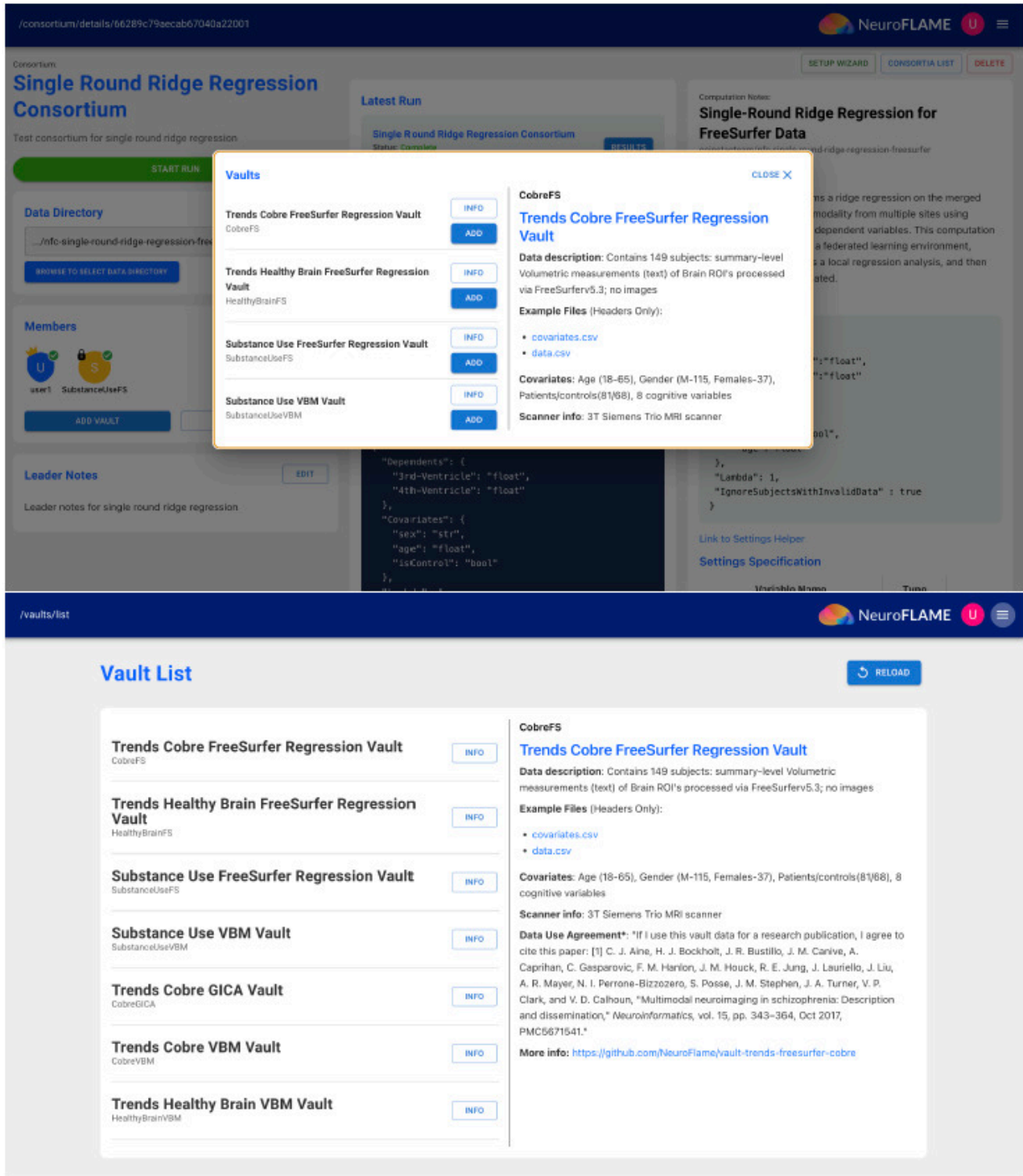


Figure 2. Screenshots from the NeuroFLAME GUI showing a list of vaults that can be added to a federated computation. A user attempting to access these vaults can choose from a list or narrow their selection based on computation-type, data-set, or other metadata. Any user not provided access to a given vault will not be allowed to add those vaults to their federated analysis.

tural MRI. OpenNeuro enables researchers to upload, download, and reuse datasets.⁴

NeuroHub is a Canadian neuroinformatics platform that integrates neuroimaging, clinical, and behavioral data with high-performance computing resources. Built on the CBRAIN infrastructure, it enables researchers to store, curate, and analyze large, multimodal neuroscience datasets

through a user-friendly web interface. NeuroHub supports data sharing and advanced analytics while complying with privacy and ethics standards, facilitating collaborative, reproducible research across institutions.¹³

EBRAINS is a European digital research infrastructure developed by the Human Brain Project. It provides a suite of tools, services, and data platforms to support brain re-

search, brain-inspired computing, and medical applications. EBRAINS offers access to high-quality datasets, brain atlases, simulation environments, and computing resources, with a strong emphasis on interoperability and collaborative research in neuroscience.¹⁴

The Virtual Imaging Platform (VIP) is a web-based platform geared to providing access to clinical data. The platform is designed to support large-scale medical imaging research by providing tools for data storage, sharing, and processing in the cloud. It enables users to run complex image analysis workflows using high-performance computing resources without requiring local software installation. VIP supports reproducibility and collaboration by allowing researchers to build, share, and execute standardized pipelines on multimodal imaging datasets, and integrates with public repositories and tools to facilitate open science in medical imaging.¹⁵

XNAT (Extensible Neuroimaging Archive Toolkit) is an open-source imaging informatics platform designed to manage, share, and analyze neuroimaging and associated data. It provides robust tools for data archiving, user access control, automated pipeline execution, and integration with imaging workflows. XNAT supports DICOM and other data formats, making it widely used in clinical and research environments for large-scale studies. Its modular design and API access enable customization and integration with external tools.¹⁶

DataLad is a Python-based tool for the joint management of code, data, and their relationship, built on top of a versatile system for data logistics (git-annex) and the most popular distributed version control system (Git), to support version control, provenance tracking, and reproducible workflows for large and complex datasets. With both CLI and GUI interfaces, DataLad simplifies collaboration across institutions and computing environments, integrates with containerized pipelines for reproducible processing, and facilitates sharing data through public platforms or custom remotes securely and efficiently.¹⁷

PySyft is an open-source Python library that enables secure, privacy-preserving machine learning through techniques like federated learning, differential privacy, and encrypted computation. Built on top of popular deep learning frameworks such as PyTorch, PySyft allows data scientists to train models on decentralized data without ever accessing the raw data itself. This makes it especially useful in sensitive domains like healthcare and finance.¹⁸

Fed-BioMed is an open-source federated learning framework designed to enable privacy-preserving machine learning across healthcare institutions without sharing raw patient data. Developed by Inria and Université Côte d'Azur, it supports multiple ML frameworks like PyTorch and MONAI and employs secure protocols to protect sensitive information during collaborative model training.¹⁹

Some platforms aimed at facilitating collaborative neuroimaging include solutions such as **BrainForge**,²⁰ **NeuroBagel**²¹ and **BrainLife**.²² BrainLife and BrainForge, for example, both allow for coordinated analyses across harmoniously annotated and processed data sets uploaded from disparate sources, and NeuroBagel facilitates the dis-

covery of similarly grouped participants across sites. Although these platforms are not used for federated analysis or learning, these efforts further emphasize the desire for large-scale solutions for coordinating analyses across studies, often through the use of data and derivative organization standards such as BIDS.¹²

With our previous federated analysis platform, called COINSTAC (Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation)²³ we have demonstrated the feasibility and scalability of federated analyses across distributed neuroimaging datasets, enabling collaborative studies without the need for centralized data sharing.²⁴⁻²⁹ COINSTAC has facilitated several large-scale, federated neuroimaging studies. For example, a landmark study involving nearly 2,000 adolescents from the IMAGEN consortium across Europe used resting-state fMRI to uncover altered brain connectivity and spectral patterns linked to smoking and alcohol use.³⁰ Another multi-site voxel-based morphometry (VBM) study spanned 14 international sites, analyzed over 2,000 adolescent brain MRIs to identify structural changes associated with age, body mass index, and smoking habits.³¹ Further, a federated VBM meta-analysis involving 4,102 subjects across eight global institutions revealed shared cortical and sub-cortical similarities across schizophrenia, major depression, and autism spectrum disorder, highlighting transdiagnostic neural signatures.³² These studies underscored the need for a federated, neuroimaging-focused platform in advancing global neuroscience by facilitating privacy-preserving, reproducible analysis of diverse, geographically distributed datasets.

As outlined in section 1, the opportunities for decentralized data sharing and collaborative analysis are huge, as researchers' needs vary drastically across domains, especially in neuroscience. This leaves room for domain-specific tools that tailor to the specific needs of principal investigators. As one of the first providers of a federated analysis platform via COINSTAC, we have learnt that researchers in the neuroimaging domain often prefer tools tailored to their domain versus a one-size-fits-all solution, which could require a lot of fine-tuning. Each of the above-mentioned tools has its place in neuroscience research as the field is still evolving. **NeuroFLAME's unique offering to neuroscience researchers is its ease of setup through use of the NVFlare backend, user-friendly GUI for global collaborative projects, pre-vetted vault datasets which are ready for analysis, and the ability for users with varying statistical analysis and research skills to contribute and participate in data analysis, as it fits their project goals.**

3. METHODS

Our motivating goal in this paper has been to enable several neuroimaging research sites to leverage information from other datasets, thereby augmenting the analysis of their data via an increased sample size. While general federated analyses offer one solution, most federated frameworks require active participation and monitoring at each site during the analysis. This means that coordinating tasks can in-

Table 1. An overview of the example vaults presented in this work. These vaults cover a number of neuroimaging modalities and feature types while using regression for federated statistical analysis. Other, non-regression computations are also supported as NeuroFLAME vaults, with any federated analysis able to be implemented as a vault if desired.

Example Vault Name	Data Set	Underlying Computation(s)	Sample Size
VBM Cobre Vault	Cobre ³³	Regression/VBM	152
FreeSurfer Cobre Vault	Cobre ³³	Regression/FreeSurfer	152
NeuroMark ICA Cobre Vault	Cobre ³³	Regression/NeuroMark ICA	152
VBM Child Mind Institute Vault	CMi ³⁴	Regression/VBM	922
FreeSurfer Child Mind Institute Vault	CMi ³⁴	Regression/FreeSurfer	922
VBM Substance Use Vault	Substance Use ³⁵	Regression/VBM	270
FreeSurfer Substance Use Vault	Substance Use ³⁵	Regression/FreeSurfer	270
COBRE Multi-Shot Regression Vault	Cobre ³³	Multi-Shot Regression/FreeSurfer	152

volve significant logistical overhead that is easily impeded by common difficulties, such as language barriers, time differences, and conflicting schedules. As mentioned previously, vaults offer a solution to this coordination problem in federated analysis. Rather than relying on active participants, vaults offer a constantly available service that allows federated analysis participants to host their data in the cloud, enabling either all active users or a list of trusted collaborators to perform federated analysis tasks as if the host is actively participating. In this article, our goal is to provide concrete examples of the kinds of vaults offered by our NeuroFLAME federated analysis toolbox, and to provide an overview of how this new software environment expands on the original vaults system provided in the previous COIN-STAC software.

3.1. VAULT OVERVIEW

In general, all NeuroFLAME vaults operate in a similar fashion. First, a site wishing to provide vault data, either locally hosted or on a cloud-based service, downloads NeuroFLAME and follows the instructions in the GUI to set up the service. The site then sets the permissions on their vault to allow open access for all NeuroFLAME users, or to provide restricted access to a chosen set of collaborators (see [figure 3](#) for an overview of vault types). If necessary, the vault owner can coordinate with the NeuroFLAME team to incorporate any novel computations or other features desired for their vault (such as determining cost, hosting multiple data types, creating different pipelines, or determining unique aggregation strategies). The vault is then made available as a part of NeuroFLAME, and any approved NeuroFLAME users will be able to incorporate the vault data into their federated analysis without the need for additional coordination with the vault owner. Because the vault data is offered as a service by NeuroFLAME, the need for complicated logistical coordination between federated sites is significantly reduced, allowing for flexible federated studies to be performed. Additionally, vaults enable users with no local data to participate in analyses just by using available vault data, a feature not otherwise available in most federated analysis frameworks.

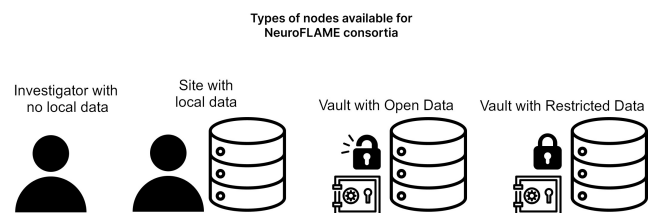


Figure 3. An overview of the four node types currently supported in NeuroFLAME. Importantly, vaults enable users with no local data to perform federated analysis tasks. Sites with local data can also combine their data with that provided by vaults. Vault owners can decide to make their data open to all NeuroFLAME users, fostering more open science efforts without significant privacy risks, and finally, vault owners may restrict access to data if desired in order to keep data private for the sake of working on publications, for example.

3.2. VAULTS UI: WORKFLOW

Users can log in to NeuroFLAME after creating an account. Once logged in, they can view all available Vaults (see [figure 2](#)), including descriptions and the analyses each Vault supports. If users require a custom analysis or wish to make a dataset available as a Vault, they can follow the integration process outlined in the “Vault Overview” section and develop custom analyses using NeuroFLAME’s boilerplate analysis creation templates. The NeuroFLAME development team provides support throughout this integration process as needed. After selecting the desired Vault data and analysis, users create a consortium, which serves as a project-based collaborative analysis workspace. Collaborators can be invited to the consortium using a secure, unique token-based authentication protocol. To facilitate the setup process, users may utilize the Vault Wizard tutorial, which guides them through configuring analyses using Vaults. Once a Vault is added to the consortium, users can modify the analysis parameters and incorporate compatible datasets either from their local machine or from other consortium members. After all datasets are mapped to the consortium, the consortium owner initiates the analysis. The

analysis pipeline runs on users' local machines where NeuroFLAME is installed, spawning Docker or Singularity containers in which the analysis code is executed. During this process, FLARE processes are initiated, and they manage the secure exchange of intermediate and summary statistics between client machines and the NeuroFLAME server, while the source data remains on users' local systems. Upon completion, each site's specific results, as well as the aggregate analysis results, are made available locally. Additionally, audit logs detailing the data transmitted between nodes are provided to users for transparency and verification.

3.3. NEUROFLAME VAULTS USE THE NVFLARE BACKEND

Although NeuroFLAME vaults use the new NVFlare backend, the core implementation is similar to the previous implementation in COINSTAC,⁸ with the main change being an effort to better unify it with that of a general federated client in the core software. Vaults use the same underlying communication and computation systems as general federated clients in NeuroFLAME (for details see³⁶), differing only in that their participation in a federated consortium does not require active participation from the host site. Thus, like general federated clients, vault data sets can be quickly used in any federated learning or analysis algorithm approved by the institution hosting the vault. In the following section, we describe some currently implemented NeuroFLAME vaults, and the data accessible to users who wish to incorporate them into federated analyses.

3.4. EXAMPLES OF IMPLEMENTED VAULTS

In this section, we provide an overview of four vaults that have been implemented within the new NeuroFLAME software. These vaults represent collaborations with multiple institutions, including the Mind Research Network, the Child Mind Institute, and the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), which coordinates neuroimaging research efforts between Georgia State University, Georgia Institute of Technology, and Emory University.

VBM COBRE Vault. The TReNDS Voxel-Based Morphometry (VBM) and FreeSurfer COBRE vaults contain preprocessed structural MRI data from 152 participants — approximately half diagnosed with schizophrenia and half healthy controls — collected as part of the Mind Research Network's COBRE study.³³ VBM computation vaults support the federated computation of a ridge-regression model with a closed-form solution in which neuroimaging data from the COBRE data set can be leveraged alongside client VBM and FreeSurfer derivatives. Accompanying the imaging data in this vault are demographic information, symptom severity scales, and cognitive measures, facilitating comprehensive model building. In addition to setting the local regularization hyperparameter, λ , clients are given the ability to build custom models using the vault data by specifying a unique choice of features and target variables from the vault to use alongside their own data.

Within the VBM vault, the data set includes gray matter images from the COBRE data set processed through a VBM pipeline³⁷ using SPM12,³⁸ encompassing reorientation, segmentation into six tissue classes, normalization to the SPM12 template, modulation, and smoothing with a $10 \times 10 \times 10$ mm Gaussian kernel. When accessing this vault data, clients are expected to perform VBM on their own data beforehand. The NeuroFLAME software facilitates this by allowing clients to perform a VBM computation using the same parameters as COBRE within the toolbox. The local VBM derivatives can then be directly piped into the downstream VBM regression model.

The vaults coordinate their federated analyses by first having each client compute summary statistics on their local data. The NeuroFLAME vault server computes the same summary statistics for the vault data, and then aggregates client statistics alongside the vault statistics to compute the closed-form solution for the ridge regression model.

FreeSurfer COBRE Vault. The FreeSurfer COBRE vault contains preprocessed structural MRI data from 152 participants — approximately half diagnosed with schizophrenia and half healthy controls — collected as part of the Mind Research Network's COBRE study.³³ The FreeSurfer vault supports the federated computation of a ridge-regression model using a single-shot averaging strategy in which sites locally optimize a regression model and the aggregator compiles these parameters to create the final model. As with the VBM vault, accompanying the imaging data in the FreeSurfer vault are demographic information, symptom severity scales, and cognitive measures, facilitating comprehensive model building. Similar to the VBM ridge regression, clients set the regularization hyperparameter, λ , and clients are given the ability to build custom models using the vault data by specifying a unique choice of features and target variables from the vault to use alongside their own data.

The FreeSurfer vault includes cortical and subcortical volumetric and surface-based measurements derived from two FreeSurfer atlases: Desikan-Killiany³⁹ and Destrieux⁴⁰ using the FreeSurfer v5.3 software.⁴¹ To generate measurements in FreeSurfer, the **recon-all** command is run on the vault data, and volumetric and surface-based measures are extracted from **aseg.stats** and **aparc.stats**. Similar to the VBM vault, clients are expected to have also performed the same FreeSurfer v5.3 analysis on their data. As with VBM, the NeuroFLAME toolbox wraps the FreeSurfer computation in order to allow clients to compute these derivatives and directly pipe them into the downstream regression model.

As with the VBM vault, the FreeSurfer vaults coordinate their federated analyses by first having each client compute local statistics on their local data set, while the NeuroFLAME vault server does the same for the vault data. The server then aggregates client statistics alongside the vault statistics to compute the closed-form solution for the ridge regression model. In contrast to sharing the closed-form solution in the VBM vaults, each local site and the vault server first perform a local linear regression to compute local parameters. The clients' parameters are then sent to the

vault server, which aggregates the β values from all clients and the vault. The vault server computes the global ridge regression parameters by averaging local parameter estimates.

NeuroMark ICA COBRE Vault. The TReNDS NeuroMark Independent Component Analysis (ICA) vault comprises resting-state functional MRI (rsfMRI) data from 179 participants, including both individuals diagnosed with schizophrenia and healthy controls, collected as part of the COBRE project.³³ The data underwent standard functional MRI preprocessing using the **SPM12** toolbox.³⁸ The first five scans were removed to allow signal equilibrium and participants' adaptation to the scanner's noise. We performed rigid body motion correction using the toolbox in SPM to correct subject head motion, followed by the slice-timing correction to account for timing differences in slice acquisition. The fMRI data were subsequently warped into the standard Montreal Neurological Institute (MNI) space using an echo-planar imaging (EPI) template and were slightly resampled to $3 \times 3 \times 3$ mm³ isotropic voxels. The resampled fMRI images were further smoothed using a Gaussian kernel with a full width at half maximum (FWHM) of 6 mm. The smoothed datasets were used in the ICA analysis.

Accompanying the imaging data in this vault are demographic information, symptom severity scales, and cognitive measures, facilitating comprehensive model building (see Mayer et al.³³ for the full list of variables collected). Similar to the other vaults, a local regression computation model is implemented, with the primary difference being the use of Functional Network Connectivity (FNC) values, which follow a NeuroMark ICA estimation of independent spatial maps and associated time series.

After preprocessing, NeuroMark ICA⁴² is used via the GIFT toolbox⁴³ to extract resting state networks from functional MRI data. Historically, implementations of data-driven network estimation techniques like ICA require researchers to manually inspect the resulting networks to determine which networks are noise and which are neurobiologically relevant. This can lead to the subjective interpretation of resulting networks and create difficulties when trying to compare or combine results from multiple sites. The NeuroMark ICA framework circumvents this for component selection in blind ICA by using templates of catered spatially independent maps from a large population⁴² to provide a spatial constraint on estimated networks so that they resemble known, canonical resting-state networks. NeuroMark ICA thus bridges the gap between fully data-driven and Atlas-based *a priori* approaches to functional network estimation, and provides a strong basis for federated computation because of the common template used between sites. Currently GIFT supports three different templates for functional MRI which include the original NeuroMark template,⁴² NeuroMark 2.2 multi-scale template with 105 components (see [figure 4](#) derived from multiple spatial scales,⁴⁴ and the NeuroMark White Matter template with 97 components (see [figure 5](#)) derived from white matter.⁴⁵

The resting-state networks estimated with NeuroMark ICA are then used to derive static functional network connectivity (sFNC) measures. First, time series are despiked and low-pass filtered at 0.15 Hz to exclude cardiac and other high-frequency artifacts. For each participant, the resulting time series for each pair of networks is used to compute correlation coefficients, forming a Component-by-Component matrix of connectivity values. The resulting connectivity values are then Fisher z-scored and can be used as features for a ridge regression model of the same form used for the other vaults.

VBM and FreeSurfer Child Mind Institute vaults. The Child Mind Institute (CMI) VBM vault comprises structural MRI data from 922 children and adolescents aged 6 to 22 years, collected as part of the Healthy Brain Network study.³⁴ This dataset includes gray matter segmentation outputs derived from a Voxel-Based Morphometry (VBM) preprocessing pipeline implemented in SPM12, using the same settings as used for the COBRE vault previously. Accompanying the imaging data are 11 variables encompassing demographic, cognitive, and substance use domains, facilitating building a comprehensive regression model.

The Child Mind Institute FreeSurfer vault contains structural MRI data from 885 children and adolescents (ages 5–21) collected as part of the Healthy Brain Network study. This vault includes cortical and subcortical volumetric measurements derived from the Desikan-Killiany FreeSurfer atlas. Researchers can also select from 11 variables spanning demographic, cognitive, and substance use domains, including total scores from Achenbach internalizing, externalizing, and p-factor scales, as well as measures like AUDIT, EHQ, IAT, and NIH Toolbox cognitive scores (e.g., Flanker and List Sorting). This vault enables the modeling of neurodevelopmental patterns in relation to mental health and behavior in youth.

Both of these vaults utilize closed-form ridge-regression similar to the other vaults for the final analysis.

VBM and FreeSurfer Substance Use vaults The Substance Use VBM vault contains gray matter segmentation data processed using the SPM VBM pipeline for 750 participants (ages 18–56; 456 male, 294 female).³⁵ Among these, 61 individuals were identified as controls, based on the number of alcohol drinking days being less than or equal to 20 days. In addition to neuroimaging, the vault includes a rich set of behavioral and clinical variables related to substance use, including AUDIT, ADS, BDI, and impulsivity scores, as well as the frequency of alcohol, cigarette, and marijuana use. Z-score normalized versions of these variables are also provided, making this vault ideal for studying the structural brain correlates of substance use behavior. This vault dataset is ready for users to run regression analysis.

The Substance use FreeSurfer vault contains structural MRI data processed using FreeSurfer for a total of 270 participants, aged 18–55 years (162 males and 108 females). Of these participants, 11 individuals were identified as controls, based on the number of alcohol drinking days of less than or equal to 20 days. This vault includes cortical and subcortical volumetric and surface-based measurements,

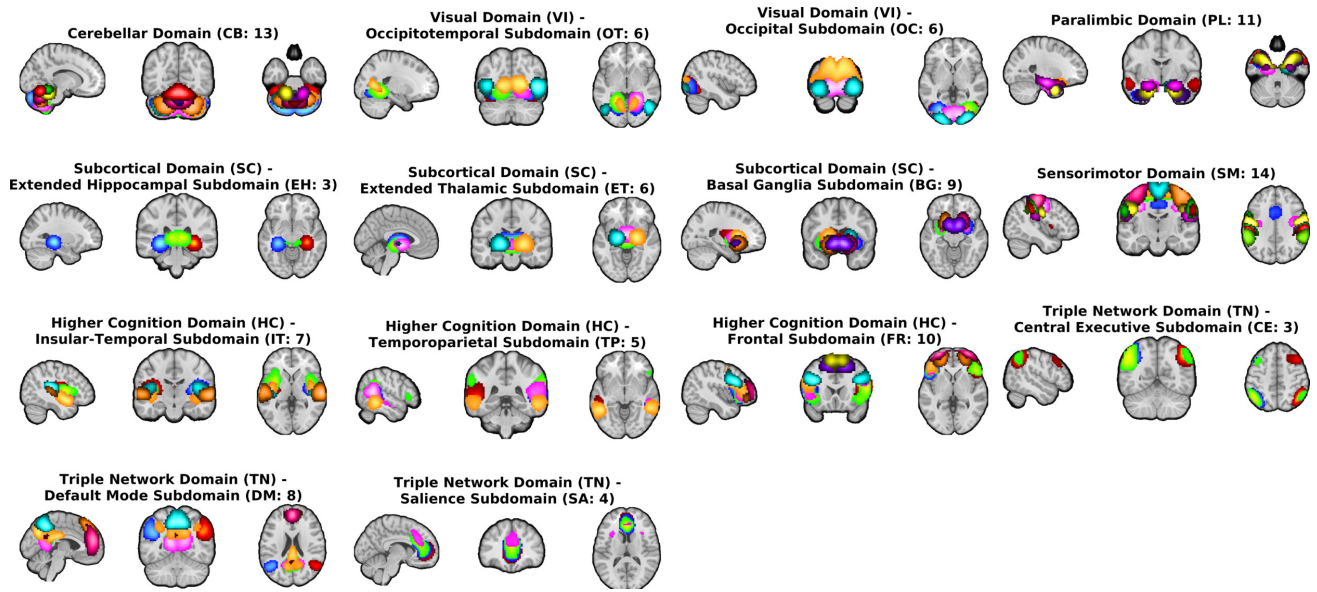


Figure 4. The 105 resting-state networks from the NeuroMark 2.2 multi-scale template.⁴⁴ The components are divided into 7 domains and 14 subdomains, which serve as a spatial guide for data-driven estimation with spatially constrained ICA.

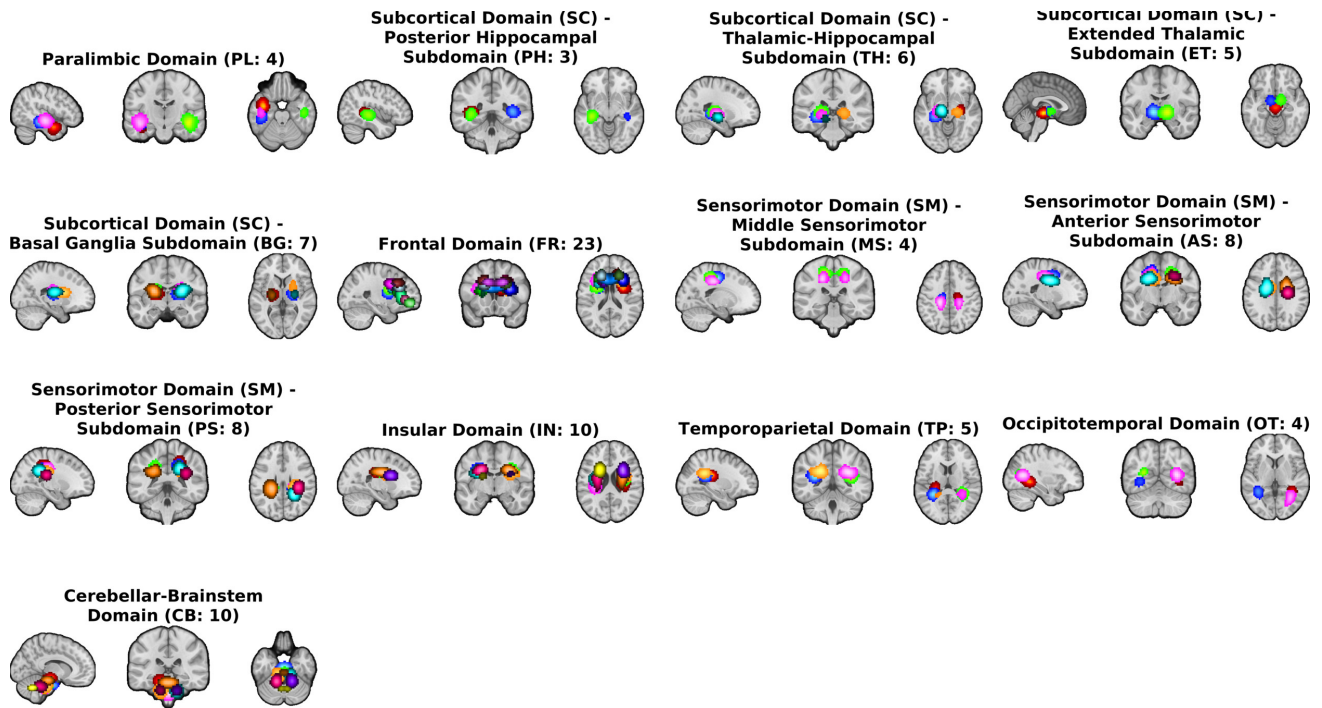


Figure 5. The 97 resting-state networks from the NeuroMark White-Matter template.⁴⁵ The components are divided into 8 domains and 6 subdomains, which serve as a spatial guide for data-driven estimation with spatially constrained ICA.

derived from FreeSurfer's comprehensive cortical reconstruction and volumetric segmentation processes. In addition to neuroimaging data, the vault includes all the behavioral and clinical variables as in the Substance Use VBM vault datasets. This vault dataset is ready for users to run regression analysis using a similar framework as described in other vaults.

COBRE MULTI-SHOT REGRESSION VAULT

The computations discussed so far represent one kind of federated analysis in which local parameters are estimated at each site, and then aggregated by computing the global or weighted mean of these parameters. In contrast to this single-shot approach, many iterative solutions to federated learning take the approach of iteratively updating parameters either by summarizing parameter gradients or param-

ters themselves. Global models are thus achieved using a federated optimization, which aims to converge on a global model identical to a model trained on the data in the pooled sample. **Because vaults can support any kind of federated computation in theory, iterative computations can also be included as vaults to be shared with collaborators.**

The COBRE multi-shot regression vault represents an example of this kind of iterative vault, and demonstrates the capacity for vaults to also support federated learning, and not just “federated analysis” in the strict sense. Using the same data set as the COBRE FreeSurfer vault, this solution allows for iterative updates to the regression parameters between the vault and participating federated collaborators, and provides similar results to the single-shot, parameter averaging in this instance. Both kinds of approaches to federated regression have been studied in the literature, and our inclusion of this vault as one of the example computations emphasizes the flexibility of vaults to support complementary computations that provide choices for consortia members accessing those data.

FUTURE VAULTS FOR OTHER FEDERATED COMPUTATIONS

The aim of presenting these example vaults is not to inherently tie NeuroFLAME vaults to a single kind of computation or data set, but to demonstrate that vaults work just like standard federated analysis solutions for neuroimaging. In principle, any federated computation that can be containerized or wrapped with Python scripts can be integrated into NeuroFLAME and offered as a vault. For example, NeuroFLAME currently supports computations for distributed, differentially private support vector machines, label noise filtering, MANCOVA, and Source-Based Morphometry, which could all be integrated into future vaults. Future vaults supporting distributed deep learning optimization, data harmonization with COMBAT, mixed effects models, parallel ICA, and ENIGMA meta-analysis will also be available on the horizon as these computations are ported from the COINSTAC software into the new NeuroFLAME environment.

3.5. ADDITIONAL PRIVACY CONSIDERATIONS FOR VAULTS

Similar to the previous COINSTAC architecture,²³ the NeuroFLAME software is built around privacy by design principles. By never sharing raw data across participating sites, it ensures that sensitive participant information remains secure and local to the originating site. Instead, only summary statistics or model parameters are exchanged during collaborative analyses, which limits the risk of data breaches or unintended re-identification. This decentralized approach aligns with modern data protection standards and allows researchers to engage in secure, multi-site studies without directly compromising participant privacy.

For sites hoping to host vaults, an initial institutional review may be required to determine specific security choices, such as which machines will host the data, which derivatives will be shared during analysis, and which partner sites

will be able to access those derivatives. A further advantage of vaults over active federated analysis participation is that vault hosting can be treated from the start as a persistent data-hosting service, rather than a dynamic collaboration, and so any logistical oversight needed can occur during the initial vault setup as well as regular review when new access needs to be provided, or changes need to be made.

4. RESULTS

In this section, we include demonstrative results for some of the vaults mentioned in the previous section. First, we provide a summary of the established results presented in previous work. We then provide a summary of new VBM regression results for the Substance use vaults, as well as new COBRE ICA results using the new multi-scale and white-matter templates.

4.1. SUMMARY OF PREVIOUS RESULTS

Our previous work demonstrated the effectiveness of decentralized data analysis via vaults in federated neuroimaging research through studies like the TReNDS VBM COBRE vault and the Child Mind Institute (CMI) VBM vault.⁸ Analyses with the TReNDS VBM COBRE vault revealed significant decreases in gray matter volume with age, sex-related differences, and schizophrenia-related reductions in insular-temporal and medial frontal regions, while combining with the MCIC dataset further improved statistical power. Using the CMI VBM vault, we observed widespread age-related gray matter volume decreases and sex-related reductions along the gray/white matter boundary. These results, consistent with prior studies, underscore how decentralized data analysis via vaults enables collaborative, privacy-preserving analyses. We utilized the TReNDS FreeSurfer COBRE vault to conduct ridge regression analyses on cortical and subcortical measurements derived from FreeSurfer atlases. The findings revealed that healthy controls exhibited significantly higher cortical thickness in temporal lobe regions compared to individuals with schizophrenia, particularly in areas such as the inferior temporal sulcus and middle temporal gyrus. These results align with established patterns of cortical thinning associated with schizophrenia and underscore the efficacy of vaults in facilitating federated neuroimaging analyses while preserving data privacy.

We used the TReNDS NeuroMark Group-ICA COBRE vault to perform univariate regression analyses on resting-state fMRI data from 179 participants. The data were pre-processed using the NeuroMark pipeline, which applies spatially constrained Independent Component Analysis (ICA) with the NeuroMark fMRI 1.0 template, encompassing 53 intrinsic connectivity networks across seven major functional domains. Our analyses revealed significant associations between age and functional network connectivity (FNC) measures, indicating age-related changes in brain connectivity patterns. Additionally, sex differences were observed in specific networks, highlighting the influence of biological factors on brain connectivity.

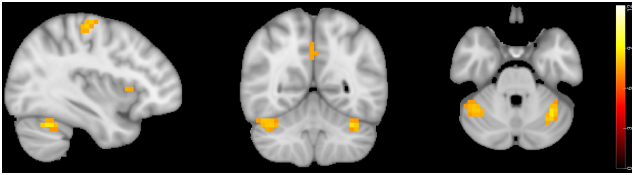


Figure 6. Significant voxels after Bonferroni correction showing the effect of alcohol consumption on gray matter volume in the Substance Use data set. These results were generated using the Substance Use VBM Vault.

More details on the results outlined above are available in our previous paper introducing vaults in the previous COINSTAC software.⁸

4.2. SUBSTANCE USE VBM REGRESSION

The Substance Use VBM vault was used to investigate the effects of alcohol consumption on gray matter volume. A significant association was found between alcohol consumption, measured by the number of drinks, and gray matter volume reduction using a regression model. The highlighted voxels in [Figure 6](#) indicate brain regions that remained significant after Bonferroni correction for multiple comparisons at a threshold of 0.05. We should note that although Bonferroni correction was used for this study, vaults are not inherently tied to any form of multiple testing correction, and FDR correction is also supported, for example. The affected areas include the bilateral cerebellum Crus I, precentral gyrus, postcentral gyrus, and anterior cingulate gyrus.

4.3. THE SUBSTANCE USE VAULT IN PRACTICE

Although a large-scale, multi-site deployment of NeuroFLAME vaults remains a goal of ongoing work, the analysis below provides an end-to-end proof of concept for the capability that distinguishes vaults from standard federated participation. In it, a remote site combines its locally held, open-access SUDMEX-CONN cocaine-use data⁴⁶ with the substance-use VBM vault and completes a full federated regression, with only intermediate statistics leaving either site ([Figure 7](#)). This demonstration is provided with no prior coordination with the host (other than initial setup of the vault as a service) because the vault responds on demand without any active involvement. This demonstration exercises exactly the persistent, coordination-free access mechanism that the vaults construct introduces, and is therefore sufficient to establish that the construct behaves as intended in the new NeuroFLAME/NVFlare environment.

The federated analysis was computed using a remote client with a copy of the NeuroFLAME software on a machine where the SUDMEX-CONN data set was downloaded. The user operated the NeuroFLAME UI (for more details see³⁶) to point the local client to the location of their local copy of the SUDMEX-CONN data set, and created a custom consortium which included the substance-use VBM vault as a participating federated node. The NeuroFLAME software both on the local client and remotely at the TReNDs center

— which as of writing is the host of the substance use vault — then aggregated the final averaged summary statistics combining the vaults analysis with the user's. The software then generated an output report and binary files containing aggregated p-values and coefficients from the regression. The negative log of the p-values are plotted for both local (no vault) and global (including vault) results in [figure 7](#).

4.4. COBRE NEUROMARK ICA WITH NEW TEMPLATES

The COBRE ICA vault was used to investigate the effects of age and sex on resting-state networks derived from the new multi-scale⁴⁴ and white-matter⁴⁵ network templates included in GIFT. Functional network connectivity (FNC) pairs of ICA time-series data were both tested for significance.

The matrix in [figure 8](#) (left) summarizes significant functional network connectivity differences from the multi-scale template⁴⁴ for sex at the subdomain level. [Figure 8](#) (right) summarizes significant differences for age at the subdomain level. For this template, the 105 resting-state networks span 7 domains and 14 subdomains: CB, Cerebellar; VI-OT, Visual Occipitotemporal; VI-OC, Visual Occipital; PL, Paralimbic; SC-EH, Subcortical Extended Hippocampal; SC-ET, Subcortical domain Extended Thalamic subdomain; SC-BG, Subcortical domain Basal Ganglia subdomain; SM, Sensorimotor; HC-IT, Higher Cognition domain Insular Temporal subdomain; HC-TP, Higher Cognition domain Temporoparietal subdomain; HC-FR, Higher Cognition domain Frontal subdomain; TN-CE, Triple Network domain Central Executive subdomain; TN-DM, Triple Network domain Default Mode subdomain; and TN-SA, Triple Network domain Salience subdomain.

For the multi-scale gray matter template,⁴⁴ in [figure 8](#) (left), we find that sex is associated with significant functional network connectivity differences between the Default mode networks (TN-DM) and the Subcortical Basal Ganglia (SC-BG) networks ($p = 0.0062$). An additional significant effect is seen between the Subcortical Extended Hippocampal (SC-EH) and Higher Cognitive Frontal (HC-FR) networks ($p = 0.0280$). For age, [figure 8](#) (right) shows significant connectivity differences within the Higher Cognitive Frontal (HC-FR) domain ($p = 0.004$), between the Paralimbic (PL) and Visual Occipitotemporal (VI-OT) networks ($p = 0.0323$), and between the Triple Network Central Executive (TN-CE) and Subcortical Extended Hippocampal (SC-EH) networks ($p = 0.0207$). The connectivity between the Triple Network Central Executive (TN-CE) network and the Visual-Occipitotemporal (VI-OT) also shows a moderately significant effect for age ($p = 0.0431$).

Similar figures to the multi-scale gray matter template are provided for the white matter template⁴⁵ in [figure 9](#) (left and right). For this template, the 97 resting-state networks span 8 domains and 6 subdomains: PL, Paralimbic; SC-PH, Subcortical domain Posterior Hippocampal subdomain; SC-ET, Subcortical domain extended thalamic subdomain; SC-BG, Subcortical domain basal ganglia subdomain; FR, Frontal; SM-MS, Sensorimotor domain middle subdomain; SM-AS, Sensorimotor domain anterior subdomain; SM-PS, Sensorimotor domain posterior subdomain; IT, in-

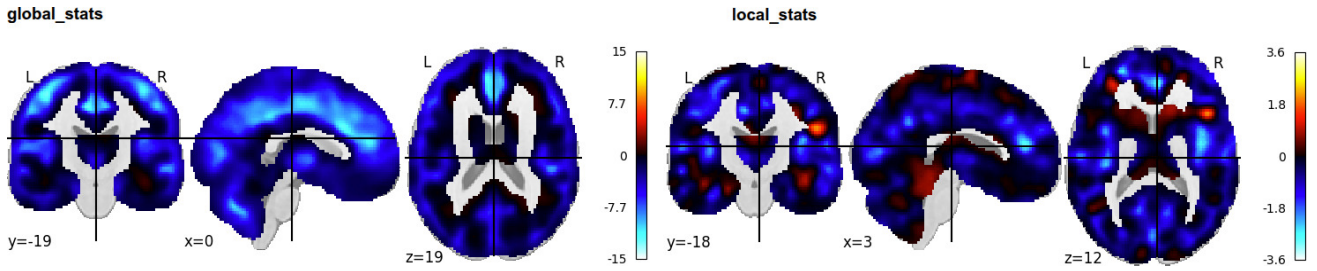


Figure 7. Federated (left) and local (right) significance scores for VBM regression using the SUDMEX-CONN data set and the substance use VBM vault. Each voxel contains the $-\log_{10}$ of the pvalue multiplied by the sign of the t -statistic.

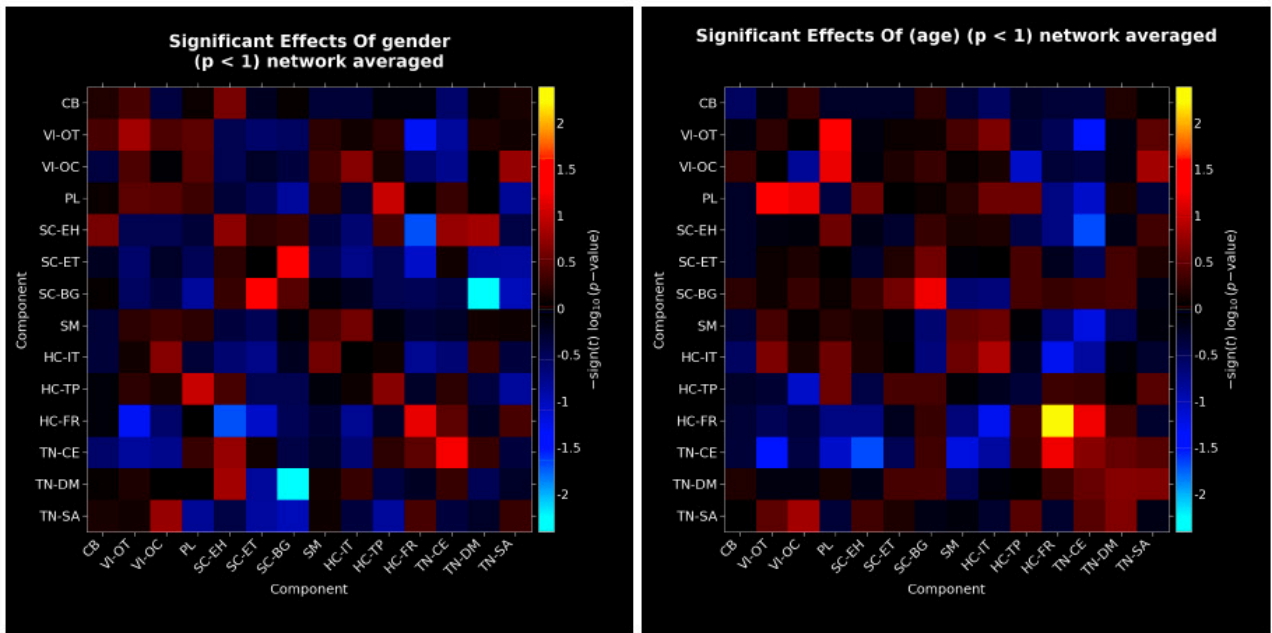


Figure 8. Effects of sex (left) and age (right) on static functional network connectivity (sFNC) at the domain level using the NeuroMark 2.2 Multi-Scale template. Each cell shows the $-\log_{10}$ of the p-value (such that values above 1.301 are significant), multiplied by the sign of the T -statistic to show the direction of the effect for a particular pair of connections using the NeuroMark2.2 template.

sular-temporal; TP, temporoparietal; OT, occipitotemporal; and CB, cerebellar.

For the white matter template,⁴⁵ in [figure 9](#) (left), we find that sex is associated with significant functional network connectivity differences within the Subcortical Extended Thalamic (SC-ET) network ($p = 0.0319$). We found no significant age effects in white matter functional network connectivity; however, connectivity between the Cerebellum (CB) and Temporoparietal (TP) networks is nearly significant ($p = 0.0535$).

5. DISCUSSION

Federated analyses offer a powerful solution to the problem of preserving privacy in large, collaborative data analyses, and our introduction of federated vaults within the new NeuroFLAME toolbox promises to address many of the logistical difficulties that come with coordinating large federated analyses with multiple sites. In this paper, we followed

up on previous work in which we introduced the concepts of vaults, providing more examples of vault computations that have been implemented and describing how the new NeuroFLAME software environment facilitates the use of these vaults.

Currently, many of the vaults use an underlying ridge regression model, due to the simplicity of averaging summary statistics for the closed-form solution; however, in principle, other, more complex models might also be able to be implemented. For example, our previous software environment contains federated algorithms for training Support Vector Machines⁴⁷ as well as more complicated Deep Learning models.⁴⁸ Recent advances in federated harmonization methods also offer immense promise for vault-based computations.^{49,50} In theory, any federated model that can be created via parameter averaging is achievable using a similar framework as with ridge regression; however, more complicated, iterative strategies that involve sharing parameter gradients between commonly initialized

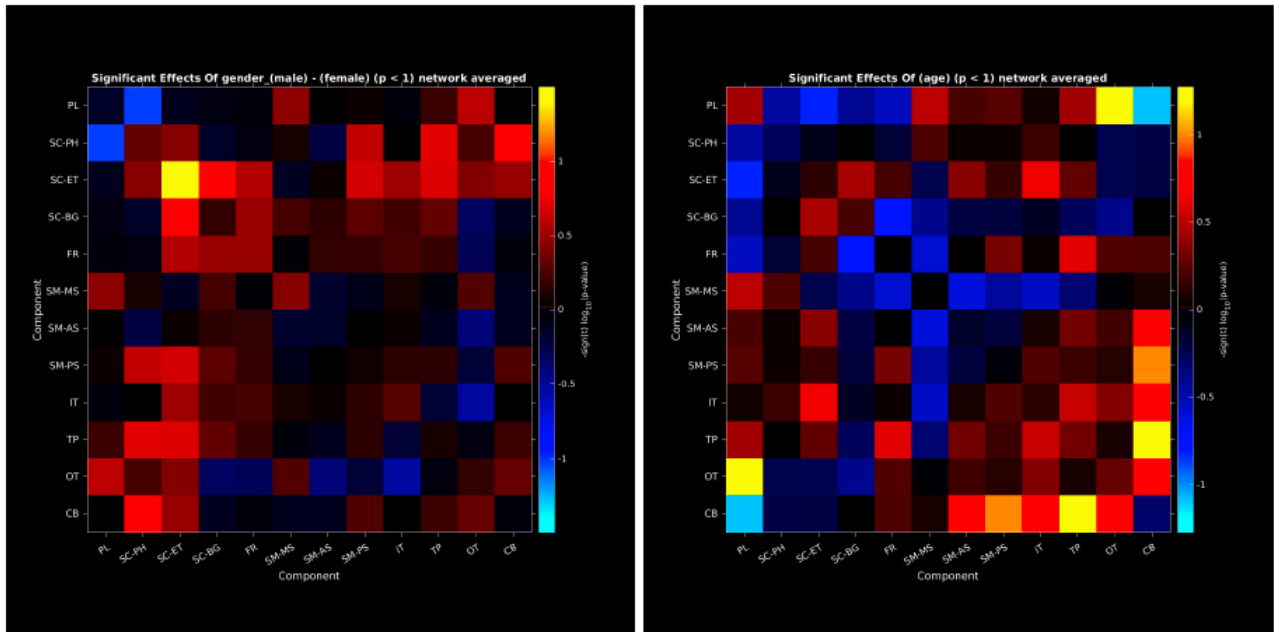


Figure 9. Effects of sex (left) and age (right) on static functional network connectivity (sFNC) at the domain level using the NeuroMark White Matter Template.

models may require some new methodological formulation to allow users full flexibility without significant computation overhead.

In addition to creating additional vault computations, we have a further goal of recruiting more research sites to contribute their data to vaults using existing computational infrastructure. We aim to offer a large set of cloud-based repositories that are shared either with the public for open use or available to selected collaborators. We would like to include multiple data sets for each computation and model type available within NeuroFLAME to allow users the flexibility to easily include multiple data sets from different sources. We specifically plan to include Schizophrenia and Substance use data sets, as well as expanding the available modalities to include electro and magnetoencephalograms (E/MEG). We plan to engage researchers across the globe through various online workshops and hackathons. Our recent Vaults workshop demonstrated substantial utility for a diverse cohort of researchers. These included graduate students and early-career scientists from the United States, Canada, Europe, and India, with interests spanning neuroimaging, federated analysis, and software development. Participants were able to run federated analyses on Vault datasets and run analyses on sample publicly accessible datasets from their personal computers. They noted that the Vaults feature lowers barriers for new researchers to explore otherwise inaccessible datasets. Many attendees expressed intentions to incorporate Vaults into their future projects using their own datasets, emphasizing the platform’s potential to democratize access to advanced neuroimaging analysis for researchers operating in resource-constrained settings. Finally, we would like to facilitate more vaults available from international sites, both in the spirit of fostering more global collaborations, but also to

help improve the generalization of neuroimaging studies to include more globe-spanning cohorts.

6. CONCLUSION

By enabling decentralized, privacy-preserving data sharing and analysis via NeuroFLAME Vaults, our platform addresses key barriers to neuroscience research, such as data sharing, privacy concerns, infrastructure limitations, and the challenges associated with integrating diverse datasets. Vaults promote the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, ensuring that research outputs are more transparent and reusable across the scientific community. As we expand our efforts to add new vaults and initiate large-scale collaborative projects, we aim to empower researchers — particularly those without access to high-performance computing infrastructure or large, heterogeneous datasets—to participate in impactful neuroimaging research. We believe our platform can serve as a catalyst for generating more inclusive, representative, and statistically robust insights into the human brain. Ultimately, this could help uncover interesting brain patterns and biomarkers by leveraging the power of large, diverse datasets analyzed in a federated and secure manner.

DATA AND CODE AVAILABILITY STATEMENT

Two publicly available datasets were analyzed in this study. These datasets can be found here: http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html, http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/sharing_neuro.html Due to the sensitive nature of the substance

use vault dataset, it has not been posted publicly, but the anonymized demographic and other clinical data used in this study are freely available for research by request. The scripts referred to in the manuscript are available on GitHub in the following repository: (<https://github.com/neuroflame>).

ACKNOWLEDGEMENTS

We thank the Vault workshop participants, in particular, Keerti Prajapati, for her feedback to improve NeuroFLAME Vaults federated analysis in reducing barriers to entry for aspiring neuroscience researchers. Her experiments using the Substance Use VBM Regression Vault also helped explore the association between alcohol consumption and gray matter volume reduction.

FUNDING SOURCES

This work was funded by the National Institutes of Health (Grants: R01DA040487, R01DA049238, and R01MH121246).

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

Submitted: June 30, 2025 CDT. Revised: January 16, 2026 CDT; May 19, 2026 CDT; July 24, 2026 CDT. Accepted: August 12, 2026 CDT. Published: September 14, 2026 CDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

REFERENCES

1. Casey BJ, Cannonier T, Conley MI, et al. The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental cognitive neuroscience*. 2018;32:43-54.
2. Miller KL, Alfaro-Almagro F, Bangarter NK, et al. Multimodal population brain imaging in the UK biobank prospective epidemiological study. *Nature neuroscience*. 2016;19(11):1523-1536.
3. Van Essen DC, Smith SM, Barch DM, et al. The WU-minn human connectome project: An overview. *Neuroimage*. 2013;80:62-79.
4. Markiewicz CJ, Gorgolewski KJ, Feingold F, et al. The OpenNeuro resource for sharing of neuroscience data. *Elife*. 2021;10:e71774.
5. Teo ZL, Jin L, Liu N, et al. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*. 2024;5(2).
6. Dwork C. Differential privacy. In: *International Colloquium on Automata, Languages, and Programming*. Springer; 2006:1-12.
7. NVIDIA. NVFlare. Published online 2025. <https://github.com/NVIDIA/NVFlare>
8. Martin D, Basodi S, Panta S, et al. Enhancing collaborative neuroimaging research: Introducing COINSTAC vaults for federated analysis and reproducibility. *Frontiers in Neuroinformatics*. 2023;17:1207721.
9. Jwa AS, Poldrack RA. The spectrum of data sharing policies in neuroimaging data repositories. *Human Brain Mapping*. 2022;43(8):2707-2721.
10. Turner MD, Gazula H, Kochunov P, et al. ENIGMA COINSTAC: Increasing neuroimaging data diversity with managed privacy. Virtual Poster presented at: APS 2020; 2020.
11. Juliano A, Davis E, Curtin G, et al. 607. Globally reduced cortical thinning in cocaine use disorder: A federated analysis of ENIGMA addiction data via COINSTAC. *Biological Psychiatry*. 2025;97(9):S348-S349.
12. Gorgolewski KJ, Alfaro-Almagro F, Auer T, et al. BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS computational biology*. 2017;13(3):e1005209.
13. Trefethen A, De Roure D, Newman D, et al. NeuroHub: A research information environment for neuroscientists. *Digital Research*. Published online 2012:10-12.
14. Schirner M, Domide L, Perdakis D, et al. Brain simulation as a cloud service: The virtual brain on EBRAINS. *NeuroImage*. 2022;251:118973.
15. Glatard T, Lartizien C, Gibaud B, et al. A virtual imaging platform for multi-modality medical image simulation. *IEEE transactions on medical imaging*. 2012;32(1):110-118.
16. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The extensible neuroimaging archive toolkit: An informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*. 2007;5:11-33.
17. Halchenko YO, Meyer K, Poldrack B, et al. DataLad: Distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software*. 2021;6(63):3262.
18. Ryffel T, Trask A, Dahl M, et al. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:181104017*. Published online 2018.
19. Silva S, Altmann A, Gutman B, Lorenzi M. Fed-biomed: A general open-source frontend framework for federated learning in healthcare. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*. Springer; 2020:201-210.
20. Verner E, Petropoulos H, Baker B, et al. Brainforge: An online data analysis platform for integrative neuroimaging acquisition, analysis, and sharing. *Concurrency and Computation: Practice and Experience*. 2023;35(18):e6855.
21. Urchs S, Dai A, Armoza J, Jahanpour A, Bhagwat N, Poline J. Neurobagel API: API for requests against graphs of neurobagel data graphs. *Published online*. Published online 2023.
22. Hayashi S, Caron BA, Heinsfeld AS, et al. Brainlife. Io: A decentralized and open-source cloud platform to support neuroscience research. *Nature methods*. 2024;21(5):809-813.

23. Plis SM, Sarwate AD, Wood D, et al. COINSTAC: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Frontiers in neuroscience*. 2016;10:365.
24. Baker BT, Abrol A, Silva RF, et al. Decentralized temporal independent component analysis: Leveraging fMRI data in collaborative settings. *NeuroImage*. 2019;186:557-569.
25. Baker BT, Damaraju E, Silva RF, Plis SM, Calhoun VD. Decentralized dynamic functional network connectivity: State analysis in collaborative settings. *Human brain mapping*. 2020;41(11):2909-2925.
26. Gazula H, Baker BT, Damaraju E, et al. Decentralized analysis of brain imaging data: Voxel-based morphometry and dynamic functional network connectivity. *Frontiers in neuroinformatics*. 2018;12:55.
27. Lewis N, Gazula H, Plis SM, Calhoun VD. Decentralized distribution-sampled classification models with application to brain imaging. *Journal of neuroscience methods*. 2020;329:108418.
28. Basodi S, Raja R, Ray B, et al. Decentralized brain age estimation using mri data. *Neuroinformatics*. 2022;20(4):981-990.
29. Basodi S, Raja R, Gazula H, et al. Decentralized mixed effects modeling in COINSTAC. *Neuroinformatics*. 2024;22(2):163-175.
30. Gazula H, Rootes-Murdy K, Holla B, et al. Federated analysis in COINSTAC reveals functional network connectivity and spectral links to smoking and alcohol consumption in nearly 2,000 adolescent brains. *Neuroinformatics*. 2023;21(2):287-301.
31. Gazula H, Holla B, Zhang Z, et al. Decentralized multisite VBM analysis during adolescence shows structural changes linked to age, body mass index, and smoking: A COINSTAC analysis. *Neuroinformatics*. Published online 2021:1-14.
32. Rootes-Murdy K, Panta S, Kelly R, et al. Cortical similarities in psychiatric and mood disorders identified in federated VBM analysis via COINSTAC. *Patterns*. 2024;5(7).
33. Mayer AR, Ruhl D, Merideth F, et al. Functional imaging of the hemodynamic sensory gating response in schizophrenia. *Human brain mapping*. 2013;34(9):2302-2312.
34. Alexander LM, Escalera J, Ai L, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*. 2017;4(1):1-26.
35. Vergara VM, Weiland BJ, Hutchison KE, Calhoun VD. The impact of combinations of alcohol, nicotine, and cannabis on dynamic brain connectivity. *Neuropsychopharmacology*. 2018;43(4):877-890.
36. Martin D, Panta S, Kelly R, et al. NeuroFLAME: A scalable, privacy-preserving federated framework for secure, reproducible, and multi-site neuroimaging analysis. *bioRxiv*. Published online 2026:2026-04.
37. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. *Neuroimage*. 2000;11(6):805-821.
38. Ashburner J, Barnes G, Chen CC, et al. SPM12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*. 2014;2464(4):53.
39. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 2006;31(3):968-980.
40. Destrieux C, Fischl B, Dale A, Halgren E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*. 2010;53(1):1-15.
41. Fischl B. Freesurfer. *Neuroimage*. 2012;62(2):774-781.
42. Du Y, Fu Z, Sui J, et al. Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders. *NeuroImage: Clinical*. 2020;28:102375.
43. Rachakonda S, Egolf E, Correa N, Calhoun V. Group ICA of fMRI toolbox (GIFT) manual. *Dostupnez [cit 2011-11-5]*. Published online 2007.
44. Iraj A, Fu Z, Faghiri A, et al. Identifying canonical and replicable multi-scale intrinsic connectivity networks in 100k+ resting-state fMRI datasets. *Human brain mapping*. 2023;44(17):5729-5748.
45. Itkyl VS, Iraj A, Jensen K, et al. Evidence for white matter intrinsic connectivity networks at rest and during a task: A large-scale study and templates. *bioRxiv*. Published online 2024:2024-12.
46. Garza-Villarreal EA, Olvera JGG, Balducci T, Valdez DA, Valencia A, Rasgado J. SUDMEX_CONN: The mexican dataset of cocaine use disorder patients. *OpenNeuro[Dataset] doi*. 2021;10.
47. Ming J, Verner E, Sarwate A, et al. COINSTAC: Decentralizing the future of brain imaging analysis. *F1000Research*. 2017;6:1512.

48. Lewis N, Plis S, Calhoun V. Cooperative learning: Decentralized data neural network. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE; 2017:324-331.

49. Bostami B, Hillary FG, Horn HJ van der, Naalt J van der, Calhoun VD, Vergara VM. A decentralized ComBat algorithm and applications to functional network connectivity. *Frontiers in Neurology*. 2022;13:826734.

50. Basodi S, Ellis CA, Bostami B, Hillary FG, Panta S, Calhoun VD. Decentralized ComBat-mega for harmonizing FNC data in coinstac. In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2025:1-4.