

# ScholarWorks@GSU

## Sharing Privacy-sensitive Access to Neuroimaging and Genetics Data: A Review and Preliminary Validation

Authors	Sarwate, Anand D.;Plis, Sergey M.;Turner, Jessica;Arbabshirani, Mohammad R.;Calhoun, Vince D.
Citation	Sarwate AD, Plis SM, Turner JA, Arbabshirani MR and Calhoun VD (2014) Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. Front. Neuroinform. 8:35. doi: <a href="https://doi.org/10.3389/fninf.2014.00035">https://doi.org/10.3389/fninf.2014.00035</a>
DOI	<a href="https://doi.org/10.3389/fninf.2014.00035">https://doi.org/10.3389/fninf.2014.00035</a>
Download date	2026-04-10 22:57:41
Link to Item	<a href="https://hdl.handle.net/20.500.14694/11507">https://hdl.handle.net/20.500.14694/11507</a>



# Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation

Anand D. Sarwate<sup>1</sup>, Sergey M. Plis<sup>2</sup>, Jessica A. Turner<sup>2,3</sup>, Mohammad R. Arbabshirani<sup>2,4</sup> and Vince D. Calhoun<sup>2,4\*</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

<sup>2</sup> Mind Research Network, Albuquerque, NM, USA

<sup>3</sup> Department of Psychology and Neuroscience Institute, Georgia State University, Atlanta, GA, USA

<sup>4</sup> Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, USA

## Edited by:

Xi Cheng, Lieber Institute for Brain Development, USA

## Reviewed by:

Adam Smith, Pennsylvania State University, USA

Sean Randall, Curtin University, Australia

## \*Correspondence:

Vince D. Calhoun, The Mind Research Network, 1101 Yale Blvd. NE, Albuquerque, NM 87106, USA  
e-mail: vcalhoun@mrn.org

The growth of data sharing initiatives for neuroimaging and genomics represents an exciting opportunity to confront the “small *N*” problem that plagues contemporary neuroimaging studies while further understanding the role genetic markers play in the function of the brain. When it is possible, open data sharing provides the most benefits. However, some data cannot be shared at all due to privacy concerns and/or risk of re-identification. Sharing other data sets is hampered by the proliferation of complex data use agreements (DUAs) which preclude truly automated data mining. These DUAs arise because of concerns about the privacy and confidentiality for subjects; though many do permit direct access to data, they often require a cumbersome approval process that can take months. An alternative approach is to only share data derivatives such as statistical summaries—the challenges here are to reformulate computational methods to quantify the privacy risks associated with sharing the results of those computations. For example, a derived map of gray matter is often as identifiable as a fingerprint. Thus alternative approaches to accessing data are needed. This paper reviews the relevant literature on differential privacy, a framework for measuring and tracking privacy loss in these settings, and demonstrates the feasibility of using this framework to calculate statistics on data distributed at many sites while still providing privacy.

**Keywords:** collaborative research, data sharing, privacy, data integration, neuroimaging

## 1. INTRODUCTION

Neuroimaging data has been the subject of many data sharing efforts, from planned large-scale collaborations such as the Alzheimers Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008) and functional biomedical informatics research network (FBIRN) (Potkin and Ford, 2009) (among others) to less-formalized operations such as openfmri.org (Poldrack et al., 2013) and the grass roots functional connectomes project (FCP) with its international extension (INDI) (Mennes et al., 2013). The Frontiers in Neuroinformatics special issue on “Electronic Data Capture, Representation, and Applications in Neuroimaging” in 2012 Turner and Van Horn (2012) included a number of papers on neuroimaging data management systems, several of which provide the research community some access to their data. In many cases, an investigator must agree to a data usage agreements (DUA): they specify who they are, what elements of the data they want, and often what they are planning to do with it. The researcher must agree to abide by arrangements such as not attempting to re-identify the subjects, not re-sharing the data, not developing a commercial application off the data, and so on. These DUAs may be as simple as a one page electronic questionnaire for contact purposes, or a full multi-page form that requires committee review, institutional official review and signatures being faxed back and forth.

The 2012 publication by members of the INCF Task Force on Neuroimaging Datasharing (Poline et al., 2012), specifically on neuroimaging data sharing, reiterated that data should be shared to improve scientific reproducibility and accelerate progress through data re-use. However, the barriers to data sharing that they identified included the well-known problems of motivation (both the ability to get credit for the data collected, as well as the fear of getting “scooped”) ethical and legal issues, and technical or administrative issues. In many cases, motivation is less of an issue than are the perceived legal and technical issues in keeping an investigator from sharing their data. The perceived legal issues regarding privacy and confidentiality, and protecting the trust that the subject has when they give their time and effort to participate in a study, are what lead to multi-page DUAs.

Neuroimaging is not the only data type whose sharing is hampered by these privacy concerns. Genetic data is perhaps the most contentious to share; the eMERGE consortium worked through a number of issues with large-scale sharing of genetic data, including the usual administrative burdens and ethical concerns (McGuire et al., 2011), and the five sites of the consortium identified numerous inconsistencies across institutional policies due to concerns about ethical and legal protections. It is often easy to re-identify individuals from genetic data; one publication showing re-identification of individuals is even possible from pooled data (Homer et al., 2008),

prompting the NIH to remove data from a public repository (Couzin, 2008). Despite the existence of more sophisticated re-identification attacks (e.g., Schadt et al., 2012), the NIH has not responded by removing the data. One of the most recent efforts re-identified subjects through combining DNA sequences with publicly available, recreational genealogy databases (Gymrek et al., 2013). These publicized privacy breaches make patients rightly concerned about their identifiable health information being shared with unknown parties.

This leads to basically three categories of data that will never be made publicly available for easy access: (1) data that are non-shareable due to obvious re-identification concerns, such as extreme age of the subject or a zip code/disease combination that makes re-identification simple; (2) data that are non-shareable due to more complicated or less obvious concerns, such as genetic data or other data which may be re-identifiable in conjunction with other data not under the investigator's control; and (3) data that are non-shareable due to the local institutional review boards (IRBs) rules or other administrative decisions (e.g., stakeholders in the data collection not allowing sharing). For example, even with broad consent to share the data acquired at the time of data collection, some of the eMERGE sites were required to re-contact the subjects and re-consent prior to sharing within the eMERGE consortium, which can be a permanent show-stopper for some datasets (Ludman et al., 2010).

The first two data types may be shared with an appropriate DUA. But this does not guarantee "easy access;" it can slow down or even prevent research. This is particularly onerous when it is not known if the data being requested are actually useable for the particular analysis the data requestor is planning. For example, it may be impossible to tell how many subjects fit a particular set of criteria without getting access to the full data first (Vinterbo et al., 2012). It is markedly problematic to spend weeks, months, or even years waiting for access to a dataset, only to find out that of the several hundred subjects involved, only a few had usable combinations of data of sufficient quality necessary for one's analysis.

Problems with DUAs only become worse when trying to access data from multiple sites. Because each DUA is different, the administrative burden rapidly becomes unmanageable. In order to enable analyses across multiple sites, one successful approach is to share data derivatives. For example, the ENIGMA consortia pooled together data from many hundreds of local sites and thousands of subjects by providing analysis scripts to local sites and centrally collecting only the output of these scripts (Hilbar et al., 2013). Another example is DataSHIELD (Wolfson et al., 2010), which also uses shared summary measures to perform pooled analysis. These systems are good starting points, but they neither quantify privacy nor provide any guarantees against re-identification. In addition, summary measures are restricted to those that can be computed independently of other data. An analysis using ENIGMA cannot iterate among sites to compute results informed by the data as a whole. However, by allowing data holders to maintain control over access, such an approach does allow for more privacy protections at the cost of additional labor in implementing and updating a distributed architecture.

The ENIGMA approach is consistent with the *differential privacy* framework (Dwork et al., 2006), a strong notion of privacy which measures the risk of sharing the results of computations on private data. This quantification allows data holders to track overall risk, thereby allowing local sites to "opt-in" to analyses based on their own privacy concerns. However, in the differential privacy model, the computation is *randomized*—algorithms introduce noise to protect privacy, thereby making the computation less accurate. However, if protecting privacy permits sharing data derivatives, then aggregating private computations across many sites may lead to a benefit; even though each local computation is less accurate (to protect privacy), the "large N" benefit from many sites allowing access will still result in a more accurate computation.

The system we envision is a research consortium in which sites allow differentially-private computations on their data without requiring an individual DUA for each site. The data stays at each site, but the private data derivatives can be exchanged and aggregated to achieve better performance. In this paper we survey some of the relevant literature on differential privacy to clarify if and how it could help provide useful privacy protections in conjunction with distributed statistical analyses of neuroimaging data. The default situation is no data sharing; each site can only learn from its own data. We performed an experiment on neuroimages from a study to see if we could predict patients with schizophrenia from healthy control subjects. Protecting privacy permits a pooled analysis; without the privacy protections, each site would have to use its own data to learn a predictor. Our experiments show that by gathering differentially private classifiers learned from multiple sites, an aggregator can create a classifier that significantly outperforms that which could be learned at a single site. This demonstrates the potential of differential privacy: sharing access to data derivatives (the classifiers) improves overall accuracy.

Many important research questions can be answered by the kind of large-scale neuroinformatics analyses that we envision.

- Regression is a fundamental statistical task. Regressing covariates such as age, diagnosis status, or response to a treatment against structure and function in certain brain regions (voxels in an image) is simple but can lead to important findings. For example, in examining the ability to aggregate structural imaging across different datasets (Fennema-Notestine et al., 2007) used the regression of age against brain volumes as a validity test. Age also affects resting state measures, as Allen et al. (2011) demonstrated on an aggregated dataset of 603 healthy subjects combined across multiple studies within an individual institution that had a commitment to data sharing and had minimal concerns regarding re-identification of the data. In that study, because privacy and confidentiality requirements that limited access to the full data, the logistics of extracting and organizing the data took the better part of a year (personal communication from the authors). In such a setting, asking a quick question such as whether age interacts with brain structure differently in healthy patients versus patients with a rare disorder would be impossible without submitting the project for IRB approval. This process can take months or even years and cost hundreds of dollars, whereas the analysis takes less than a day and may

produce negative findings. We need a framework that facilitates access to data on the fly for such straightforward but fundamental analyses.

- The re-use of genetic data has been facilitated by dbGAP, NIH's repository for sharing genome-wide scan datasets, gene expression datasets, methylation datasets, and other genomic measures. The data need to be easily accessible for combined analysis for identification or confirmation of risk genes. The success of the Psychiatric Genomic Consortium in finding confirmed risk genes of schizophrenia after almost 5 years of aggregating datasets supports these goals of making every dataset re-usable (Ripke et al., 2013). While dbGAP has been a resounding success, it has its drawbacks. Finding the data can be a bit daunting, as often phenotype data is made available separately from the genetic data. For example, the PREDICT-HD Huntington's disease study rolled out a year before the genetic data. DbGAP's sharing requirements are driven by the need to ensure the data are handled appropriately and the subjects' confidentiality and privacy are protected; requesting a dataset entails both the PI and their institutional official signing an agreement as well as a review by the study designate. This process must be completed prior to access being granted or denied. As before, this precludes any exploratory analyses to identify particular needs, such as determining how many subjects have the all the required phenotype measures.
- The success of multimodal data integration in the analysis of brain structure/function (Plis et al., 2010; Bießmann et al., 2011; Bridwell et al., 2013; Schelenz et al., 2013), imaging/genetics (Liu et al., 2012; Chen et al., 2013; van Erp et al., 2013), and EEG/fMRI (Bridwell et al., 2013; Schelenz et al., 2013) shows that with enough data, we can go further than simple univariate linear models. For example, we can try to find combinations of features which predict the development of a disorder, response to various treatments, or relapse. With more limited data there has been some success in reproducing diagnostic classifications (Arbabshirani et al., 2013; Deshpande et al., 2013), and identifying coherent subgroupings within disorders which may have different genetic underpinnings (Girirajan et al., 2013). With combinations of imaging, genetic, and clinical profiles from thousands of subjects across autism, schizophrenia, and bipolar disorder, for example, we could aim to identify more clearly the areas of overlap and distinction, and what combinations of both static features and dynamic trajectories in the feature space identify clinically relevant clusters of subjects who may be symptomatically ambiguous.

## 2. PRIVACY MODELS AND DIFFERENTIAL PRIVACY

There are several different conceptual approaches to defining privacy in scenarios involving data sharing and computation. One approach is to create *de-identified* data; these methods take a database of records corresponding to individuals and create a *sanitized database* for use by the public or another party. Such approaches are used in official statistics and other settings—a survey of different privacy models can be found in Fung et al. (2010), and a survey of privacy technologies in a medical informatics context in Jiang et al. (2013). These approaches differ in how they

define privacy and what guarantees they make with respect to this definition. For example, *k*-anonymity (Sweeney, 2002) quantifies privacy for a particular individual *i* with data  $x_i$  (for example, age and zip code) in terms of the number of other individuals whose data is also equal to  $x_i$ . Algorithms for guaranteeing *k*-anonymity manipulate data values (e.g., by reporting age ranges instead of exact ages) to enforce that each individual's record is identical to at least *k* other individuals.

A different conceptual approach to defining privacy is to try and quantify the change in the risk of re-identification as a result of publishing a function of the data. This differs from data sanitizing methods in two important respects. Firstly, privacy is a property of an algorithm operating on the data, rather than a property of the sanitized data—this is the difference between *semantic* and *syntactic* privacy. Secondly, it can be applied to systems which do not share data itself but instead share data derivatives (functions of the data). The recently proposed  $\epsilon$ -differential privacy model (Dwork et al., 2006) quantifies privacy in terms of risk; it bounds the likelihood that someone can re-infer the data of an individual. Algorithms that guarantee differential privacy are *randomized*—they manipulate the data values (e.g., by adding noise) to bound the risk.

Finally, some authors define privacy in terms of data security and say that a data sharing system is private if it satisfies certain cryptographic properties. The most common of these models is secure multiparty computation (SMC) (Lindell and Pinkas, 2009), in which multiple parties can collaborate to compute a function of their data without leaking information about their private data to others. The guarantees are cryptographic in nature, and do not assess the re-inference or re-identification problem. For example, in a protocol to compute the maximum element across all parties, a successful execution would reveal the maximum. A secondary issue is developing practical systems to work on neuroinformatics data. Some progress has been made in this direction (Sadeghi et al., 2010; Huang et al., 2011; Nikolaenko et al., 2013), and it is conceivable that in a few years SMC will be implemented in real distributed systems.

### 2.1. PRIVACY TECHNOLOGIES FOR DATA SHARING

As discussed earlier, there are many scenarios in which sharing raw data is either difficult or impossible—strict DUAs, obvious re-identification issues, difficulties in assessing re-identifiability, and IRB or other policy rules. Similar privacy challenges exist in the secondary use of clinical data (National Research Council, 1997). In many medical research contexts, there has been a shift toward sharing *anonymized* data. The Health Insurance Portability and Accountability Act (HIPAA) privacy rule (45 CFR Part 160 and Subparts A and E of Part 164) allows the sharing of data as long as the data is de-identified. However, many approaches to anonymizing or “sanitizing” data sets (Sweeney, 2002; Li et al., 2007; Machanavajjhala et al., 2007; Xiao and Tao, 2007; Malin, 2008) are subject to attacks (Sweeney, 1997; Ganta et al., 2008; Narayanan and Shmatikov, 2008) that use public data to compromise privacy.

When data sharing itself is precluded, methods such as *k*-anonymity (Sweeney, 2002), *l*-diversity (Machanavajjhala et al., 2007), *t*-closeness (Li et al., 2007), and *m*-invariance (Xiao and

Tao, 2007) are no longer appropriate, since they deal with constructing private or sanitized versions of the data itself. In such situations we would want to construct data access *systems* in which data holders do not share the data itself but instead provide an interface to the data that allows certain pre-specified computations to be performed on that data. The data holder can then specify the granularity of access it is willing to grant subject to its policy constraints.

In this model of *interactive data access*, the software that controls the interface to the raw data acts as a “curator” that screens queries from outsiders. Each data holder can then specify the level of access which it will provide to outsiders. For example, a medical center may allow researchers to access summaries of clinical data for the purposes of exploratory analysis; a researcher can assess the feasibility of doing a study using existing records and then file a proposal with the IRB to access the real data (Murphy and Chueh, 2002; Murphy et al., 2006; Lowe et al., 2009; Vinterbo et al., 2012). In the neuroinformatics context, data holders may allow outside users to receive a histogram of average activity levels for regions of a certain size.

Being able to track the privacy risks in such an interactive system allows data holders to match access levels with local policy constraints. The key to privacy tracking is *quantification*—for each query or access to the data, a certain amount of information is “leaked” about the underlying data. With a sufficient number of queries it is theoretically possible to reconstruct the data (Dinur and Nissim, 2003), so the system should be designed to mitigate this threat and allow the data holders to “retire” data which has been accessed too many times.

## 2.2. DIFFERENTIAL PRIVACY

A user of the database containing private information may wish to apply a *query* or algorithm to the data. For example, they may wish to know the histogram of activity levels in a certain brain region for patients with a specified mutation. Because the answer to this query is of much lower dimension than a record in the database, it is tempting to regard disclosing the answer as not incurring a privacy risk. An important observation of Dinur and Nissim (2003) was that an adversary posing such queries may be able to reconstruct the entire database from the answers to multiple simple queries. The *differential privacy* model was introduced shortly thereafter, and has been adopted widely in the machine learning and data mining communities. The survey by Dwork and Smith (2009) covers much of the earlier theoretical work, and Sarwate and Chaudhuri (2013) review some works relevant to signal processing and machine learning. In the basic model, the database is modeled as a collection of  $N$  individuals’ data records  $\mathcal{D} = (x_1, x_2, \dots, x_N)$ , where  $x_j$  is the data for individual  $j$ . For example,  $x_j$  may be the MRI data associated to individual  $j$  together with information about mutations in certain genes for that individual.

An even simpler example is to estimate the mean activity in a certain region, so each  $x_j$  is simply a scalar which represented the measured activity of individual  $j$ . Let us call this desired algorithm  $\text{Alg}$ . Without any privacy constraint, the data curator would simply apply  $\text{Alg}$  to the data  $\mathcal{D}$  to produce an output  $h = \text{Alg}(\mathcal{D})$ . However, in many cases the output  $h$  could

compromise the privacy of the data and unfettered queries could lead to reidentification of an individual.

Under differential privacy, the curator applies an approximation  $\text{PrivAlg}$  to the data instead of  $\text{Alg}$ . The approximation  $\text{PrivAlg}$  is *randomized*—the randomness of the algorithm ensures that an observer of the output will have a difficult time re-identifying any individual in the database. More formally,  $\text{PrivAlg}(\cdot)$  provides  $\epsilon$ -differential privacy if for any subset of outputs  $\mathcal{S}$ ,

$$\mathbb{P}(\text{PrivAlg}(\mathcal{D}) \in \mathcal{S}) \leq e^\epsilon \cdot \mathbb{P}(\text{PrivAlg}(\mathcal{D}') \in \mathcal{S}) \quad (1)$$

for any two databases  $\mathcal{D}$  and  $\mathcal{D}'$  differing in a single individual. Here  $\mathbb{P}(\cdot)$  is the probability over the randomness in the algorithm. It provides  $(\epsilon, \delta)$ -differential privacy if

$$\mathbb{P}(\text{PrivAlg}(\mathcal{D}) \in \mathcal{S}) \leq e^\epsilon \mathbb{P}(\text{PrivAlg}(\mathcal{D}') \in \mathcal{S}) + \delta. \quad (2)$$

The guarantee that differential privacy makes is that the distribution of the output of  $\text{PrivAlg}$  does not change too much, regardless of whether any individual  $x_j$  is in the database or not. In particular, an adversary observing the output of  $\text{PrivAlg}$  and knowing all of the data of individuals in  $\mathcal{D} \cap \mathcal{D}'$  common to both  $\mathcal{D}$  and  $\mathcal{D}'$  will still be uncertain of the remaining individual’s data. Since this holds for any two databases which differ in one data point, each individual in the database is guaranteed of this protection. More specifically, the parameters  $\epsilon$  and  $\delta$  control the tradeoff between the false-alarm (Type I) and missed-detection (Type II) errors for an adversary trying to make a test between  $\mathcal{D}$  and  $\mathcal{D}'$  (see Oh and Viswanath, 2013 for a discussion).

Returning to our example of estimating the mean, the desired algorithm  $\text{Alg}$  is simply the sample mean of the  $m$  data points, so  $\text{Alg}(\mathcal{D}) = \frac{1}{m} \sum_{j=1}^m x_j$ . The algorithm  $\text{Alg}$  itself does not provide privacy because output is deterministic: the distribution of  $\text{Alg}(\mathcal{D})$  is a point mass exactly at the average. If we change one data point to form, say  $\mathcal{D}' = (x_1, x_2, \dots, x_{m-1}, x'_m)$ , then  $\text{Alg}(\mathcal{D}') \neq \text{Alg}(\mathcal{D})$  and the only way Equation (1) can hold is if  $\epsilon = \infty$ . One form of a private algorithm is to add noise to the average (Dwork et al., 2006). A differentially private algorithm is  $\text{PrivAlg}(\mathcal{D}) = \frac{1}{m} \sum_{j=1}^m x_j + \frac{1}{\epsilon m} z$ , where  $z$  has a Laplace distribution with unit variance. The Laplace distribution is a popular choice, but there are many other distributions which can also guarantee differential privacy and may be better in some settings (Geng and Viswanath, 2012, 2013). For more general functions beyond averages, Gupte and Sundararajan (2010) and Ghosh et al. (2012) showed that in some cases we can find optimal mechanisms, while Nissim and Brenner (2010) show that this optimality may not be possible in general.

Although some variations on these basic definition have been proposed in the literature (Chaudhuri and Mishra, 2006; Rastogi et al., 2009; Kifer and Machanavajjhala, 2011), most of the literature focuses on  $\epsilon$ - or  $(\epsilon, \delta)$ -differential privacy. Problems that have been studied in the literature range from statistical estimation (Smith, 2011; Kifer et al., 2012; Smith and Thakurta, 2013), to cover more complex data processing algorithms such as real-time signal processing (Fan and Xiong, 2012; Le Ny and Pappas, 2012a,b), classification (Chaudhuri et al., 2011; Rubinstein et al.,

2012; Zhang et al., 2012b; Jain and Thakurta, 2014), online learning (Jain et al., 2012; Thakurta and Smith, 2013), dimensionality reduction (Hardt et al., 2012; Chaudhuri et al., 2013), graph estimation (Karwa et al., 2011; Kasiviswanathan et al., 2013), and auction design (Ghosh and Roth, 2011). The preceding citations are far from exhaustive, and new papers on differential privacy appear each month as methods and algorithms become more mature.

There are two properties of differential privacy which enable the kind of *privacy quantification* that we need in shared data-access scenarios. The first property is *post-processing invariance*: the output of an  $\epsilon$ -differentially private algorithm  $\text{PrivAlg}$  maintains the same privacy guarantee—if  $\hat{h} = \text{PrivAlg}(\mathcal{D})$ , then the output of any function  $g(\hat{h})$  applied to  $\hat{h}$  is also  $\epsilon$ -differentially private, provided  $g(\cdot)$  doesn't depend on the data. This means that once the data curator has guaranteed  $\epsilon$ -differential privacy for some computation, it need not track how the output is used in further processing. The second feature is *composition*—if we run two algorithms  $\text{PrivAlg}_1$  and  $\text{PrivAlg}_2$  on data  $\mathcal{D}$  with privacy guarantees  $\epsilon_1$  and  $\epsilon_2$ , then combined they have privacy risk at most  $\epsilon_1 + \epsilon_2$ . In some cases these composition guarantees can be improved (Dwork et al., 2010; Oh and Viswanath, 2013).

### 2.3. DIFFERENTIALLY PRIVATE ALGORITHMS

A central challenge in the use of differentially private algorithms is that by using randomization to protect privacy, the corresponding accuracy, or *utility*, of the result is diminished. We contend that the potential for a much larger sample size through data sharing makes this tradeoff worthwhile. In this section we discuss some of the differentially private methods for statistics and machine learning that have been developed in order to help balance privacy and utility in data analyses.

Differentially private algorithms have been developed for a number of important fundamental tasks in basic statistics and machine learning. Wasserman and Zhou (2010) put the differential privacy framework in a general statistical setting, and Smith (2011) studied point estimation, showing that many statistical quantities can be estimated with differential privacy with similar statistical efficiency. Duchi et al. (2012, 2013) studied a different version of *local* privacy and showed that requiring privacy essentially entails an increase in the sample size. Since differential privacy is related to the stability of estimators under changes in the data, Dwork and Lei (2009) and Lei (2011) used tools from robust statistics to design differentially private estimators. Williams and McSherry (2010) studied connections to probabilistic inference. More recently, Kifer et al. (2012) proposed methods for high-dimensional regression and Smith and Thakurta (2013) developed a novel variable selection method based on the LASSO.

One approach to designing estimators is the sample-and-aggregate (Nissim et al., 2007; Smith, 2011; Kifer et al., 2012), which uses subsampling of the data to build more robust estimators. This approach was applied to problems in sparse linear regression (Kifer et al., 2012), and in particular to analyze the LASSO (Smith and Thakurta, 2013) under the slightly weaker definition of  $(\epsilon, \delta)$ -differential privacy. There are several works which address convex optimization approaches to

statistical model selection and machine learning under differential privacy (Chaudhuri et al., 2011; Kifer et al., 2012; Rubinstein et al., 2012; Zhang et al., 2012b) that encompass popular methods such as logistic regression, support vector machines, and other machine learning methods. Practical kernel-based methods for learning with differential privacy are still in their infancy (Chaudhuri et al., 2011; Jain and Thakurta, 2013).

### 2.4. CHALLENGES FOR DIFFERENTIAL PRIVACY

In addition to the theoretical and algorithmic developments, some authors have started trying to build end-to-end differentially private analysis toolkits and platforms. The query language PINQ (McSherry, 2010) was the first tool that allowed people to write differentially-private data-analysis programs that guarantee differential privacy, and has been used to write methods for a number of tasks, including network analyses (McSherry and Mahajan, 2010). Fuzz (Reed and Pierce, 2010) is a functional programming language that also guarantees differential privacy. At the systems level, AIRAVAT (Roy et al., 2010) is a differentially private version of MapReduce and GUPT (Mohan et al., 2012) uses the sample-and-aggregate framework to run general statistical algorithms such as  $k$ -means. One of the lessons from these implementations is that building a differentially private *system* involves keeping track of every data access—each access can leak some privacy—and systems can be vulnerable to attack from adversarial queries (Haeberlen et al., 2011).

A central challenge in designing differentially private algorithms for practical systems is setting the privacy risk level  $\epsilon$ . In some cases,  $\epsilon$  must be chosen to be quite large in order to produce useful results—such a case was studied in earlier work by Machanavajjhala et al. (2008) in the context of publishing differentially private statistics about commute times. On the other side, choosing a small value of  $\epsilon$  may result in adding too much noise to allow useful analysis. To implement a real system, it is necessary to do a proper evaluation of the impact of  $\epsilon$  on the utility of the results. Ultimately, the setting of  $\epsilon$  is a policy decision that is informed by the privacy-utility tradeoff.

There are several difficulties with implementing existing methods “off the shelf” in the neuroinformatics context. Neuroimaging data is often continuous-valued. Much of the work on differential privacy has focused on discrete data, and algorithms for continuous data are still being investigated theoretically (Sarwate and Chaudhuri, 2013). In this paper we adapt existing algorithms, but there is a need to develop methods specifically designed for neuroimage analyses. In particular, images are high-dimensional signals, and differentially private version of algorithms such as PCA may perform poorly as the data dimension increases (Chaudhuri et al., 2013). Some methods do exist that exploit structural properties such as sparsity (Hardt and Roth, 2012, 2013), but there has been insufficient empirical investigation of these methods. Developing low-dimensional representations of the data (perhaps depending on the task) can help mitigate this.

Finally, neuroimaging datasets may contain few individuals. While the signal from each individual may be quite rich, the number of individuals in a single dataset may be small. Since

privacy affects the statistical efficiency of estimators, we must develop distributed algorithms that can leverage the properties of datasets at many locations while protecting the privacy of the data at each. Small sample sizes present difficulties for statistical inference without privacy—the hope is that the larger sample size from sharing will improve statistical inference despite the impact of privacy considerations. We illustrate this in the next section.

### 3. APPLYING DIFFERENTIAL PRIVACY IN NEUROINFORMATICS

In the absence of a substitute for individual DUAs, sites are left to perform statistical analyses on their own data. Our proposal is to have sites participate in consortium in which they share differentially private data derivatives, removing the need for individual DUAs. Differential privacy worsens the quality of a statistical estimate at a single site because it introduces extra noise. However, because we can share the results of differentially private computations at different sites, we can reduce the impact of the noise from privacy. This larger effective sample size can give better estimates than are available at a single site, even with privacy. We illustrate this idea with two examples. The first is a simple problem of estimating the mean from noisy samples, and the second is an example of a classification problem.

#### 3.1. ESTIMATING A MEAN

Perhaps the most fundamental statistical problem is estimating the mean of a variable. Suppose that we have  $N$  sites, each with  $m$  different samples of an unknown effect:

$$x_{i,j} = \mu + z_{i,j} \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, m, \quad (3)$$

where  $\mu$  is an unknown mean, and  $z_{i,j}$  is normally distributed noise with zero mean and unit variance. Each site can compute its local sample mean:

$$\bar{X}_i = \frac{1}{m} \sum_{j=1}^m x_{i,j} = \mu + \frac{1}{m} \sum_{j=1}^m z_{i,j}. \quad (4)$$

The sample mean  $\bar{X}_i$  is an estimate of  $\mu$  which has an error that is normally distributed with zero mean and variance  $\frac{1}{m}$ . Thus a single site can estimate  $\mu$  to within variance  $\frac{1}{m}$ . A simple  $\epsilon$ -differentially private estimate of  $\mu$  is

$$\tilde{X}_i = \frac{1}{m} \sum_{j=1}^m x_{i,j} + \frac{1}{\epsilon m} w_i, \quad (5)$$

where  $w_i$  is a Laplace random variable with unit variance. Thus a single site can make a differentially private estimate of  $\mu$  with error variance  $\frac{1}{m} + \frac{1}{(\epsilon m)^2}$ . Now turning to the  $N$  sites, we can form an overall estimate using the differentially private local estimates:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N \tilde{X}_i = \mu + \frac{1}{mN} \sum_{i=1}^N \sum_{j=1}^m x_{i,j} + \frac{1}{\epsilon mN} \sum_{i=1}^N w_i. \quad (6)$$

This is an estimate of  $\mu$  with variance  $\frac{1}{mN} + \frac{1}{(\epsilon m)^2 N}$ .

The data sharing solution results in a lower error compared to the local non-private solution whenever  $\frac{1}{m} > \frac{1}{mN} + \frac{1}{(\epsilon m)^2 N}$ , or

$$N > 1 + \frac{1}{\epsilon^2 m}.$$

As the number of sites increases, we can support additional privacy at local nodes ( $\epsilon$  can decrease) while achieving superior statistical performance over learning at a single site *without privacy*.

#### 3.2. CLASSIFICATION

We now turn to a more complicated example of differentially private classification that shows how a public data set can be enhanced by information from differentially private analyses of additional data sets. In particular, suppose there are  $N$  sites with private data and 1 site with a publicly available dataset. Suppose private site  $i$  has  $m_i$  data points  $\{(\tilde{x}_{i,j}, y_{i,j}) : j = 1, 2, \dots, m_i\}$ , where each  $\tilde{x}_{i,j} \in \mathbb{R}^d$  is a  $d$ -dimensional vector of numbers representing features of the  $j$ -th individual at site  $i$ , and  $y_{i,j} \in \{-1, 1\}$  is a label for that individual. For example, the data could be activity levels in certain voxels and the label could indicate a disease state. Each site can learn a classifier on its own local data by solving the following minimization problem.

$$\tilde{w}_i = \operatorname{argmin}_{\tilde{w} \in \mathbb{R}^d} \sum_{j=1}^{m_i} \ell(y_{i,j} \tilde{w}^\top \tilde{x}_{i,j}) + \frac{\lambda}{2} \|\tilde{w}\|^2, \quad (7)$$

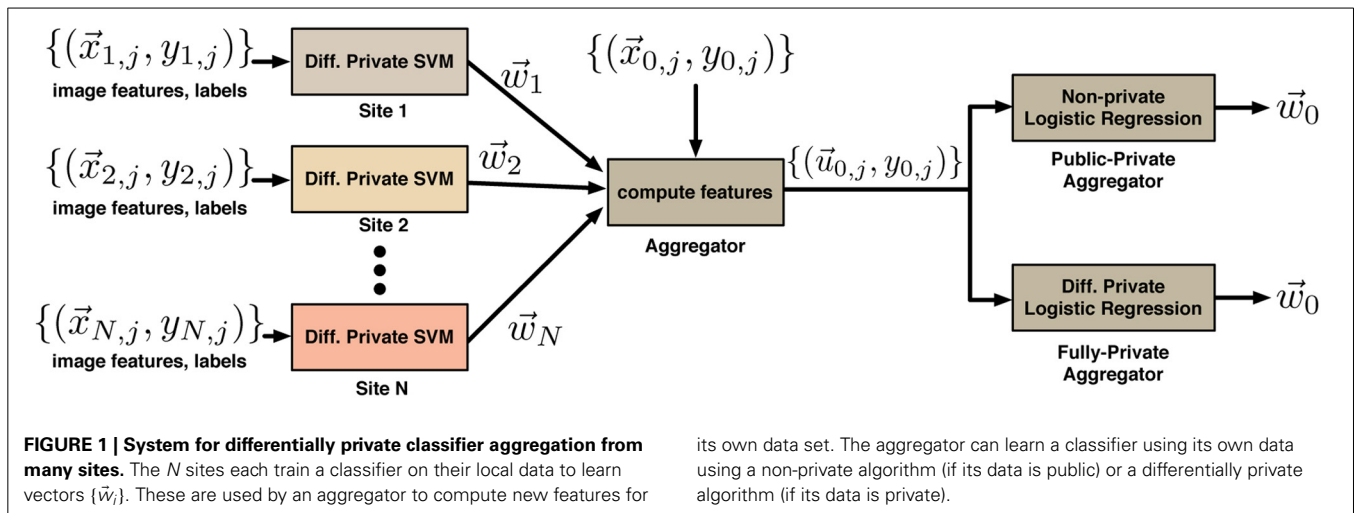
where  $\ell(\cdot)$  is a loss function. This framework includes many popular algorithms: for the support vector machine (SVM)  $\ell(z) = \max(0, 1 - z)$  and for logistic regression  $\ell(z) = \log(1 + e^{-z})$ .

Because the data at each site might be limited, they may benefit from producing differentially private versions  $\tilde{w}_i$  and then combining those with the public data to produce a better overall classifier. That is, leveraging many noisy classifiers may give better results than any  $\tilde{w}_i$  on its own. The method we propose is to train  $N$  differentially private classifiers using the objective perturbation method applied to the Huberized support vector machine (see Chaudhuri et al., 2011 for details). In this procedure, the local sites minimize a perturbed version of the classifier given in Equation (7). Let  $\tilde{w}_i$  be the differentially private classifier produced by site  $i$ .

Suppose the public data set has  $m_0$  points  $\{(\tilde{x}_{0,j}, y_{0,j}) : j = 1, 2, \dots, m_0\}$ . We compute a new data set  $\{(\tilde{u}_{0,j}, y_{0,j}) : j = 1, 2, \dots, m_0\}$  where  $\tilde{u}_{0,j}$  is an  $N$ -dimensional vector whose  $i$ -th component is equal to  $\tilde{w}_i^\top \tilde{x}_{0,j}$ . Thus  $\tilde{u}_{0,j}$  is the vector of “soft” predictions of the  $N$  differentially private classifiers produced by the private sites. The public site then uses logistic regression to train a new classifier:

$$\tilde{w}_0 = \operatorname{argmin}_{\tilde{w} \in \mathbb{R}^d} \sum_{j=1}^{m_0} \log(1 + e^{-y_{0,j} \tilde{w}^\top \tilde{u}_{0,j}}) + \frac{\lambda}{2} \|\tilde{w}\|^2. \quad (8)$$

This procedure is illustrated in **Figure 1**. The overall classification system produced by this procedure consists of the classifiers



$\{\vec{w}_i : i = 0, 1, \dots, N\}$ . To classify a new point  $\vec{x} \in \mathbb{R}^d$ , the system computes  $\vec{u} = (\vec{w}_1^\top \vec{x}, \vec{w}_2^\top \vec{x}, \dots, \vec{w}_N^\top \vec{x})$  and then predicts the label  $\hat{y} = \text{sign}(\vec{w}_0^\top \vec{u})$ . In the setting where the public site has more data, training a classifier on pairs  $(\vec{u}, \vec{x})$  could also work better.

We can distinguish between two cases here—in the *public-private* case, described above, the classifier in Equation (8) uses differentially private classifiers from each of the  $N$  sites on public data, so the overall algorithm is differentially private with respect to the private data at the  $N$  sites. In the *fully-private* case, the data at the  $(N + 1)$ -th site is also private. In this case we can replace Equation (8) with a differentially private logistic regression method (Chaudhuri et al., 2011) to obtain a classifier which is differentially private with respect to the data at all  $N + 1$  sites. Note, although we assign the role of constructing the overall two-level classifier to either the public-data site or one of the private sites in the real use-case no actual orchestrating of the process is required. It is convenient for the purposes of the demonstration (and without loss of generality) to treat a pre-selected site as an aggregator, which we do in the experiments below. **Figure 2.** can only be interpreted if we are consistent with the site that does the aggregation. However, all that needs to be done for the whole system to work is for the  $N$  (or  $N + 1$  in the fully private case) private sites compute and publish their classifiers  $\vec{w}_i$ . Then in the public data case, anyone (even entities with no data), can construct and train a classifier by simply downloading the publicly available dataset and following the above-described procedure. This could be one of the sites with the private data as well. When no public data is available the second level classifier can be only computed by one of the private-data sites (or each one of them) and later published online to be useful even for entities with insufficient data. In both cases, the final classifier (or classifiers) is based on a larger data pool that is available to any single site.

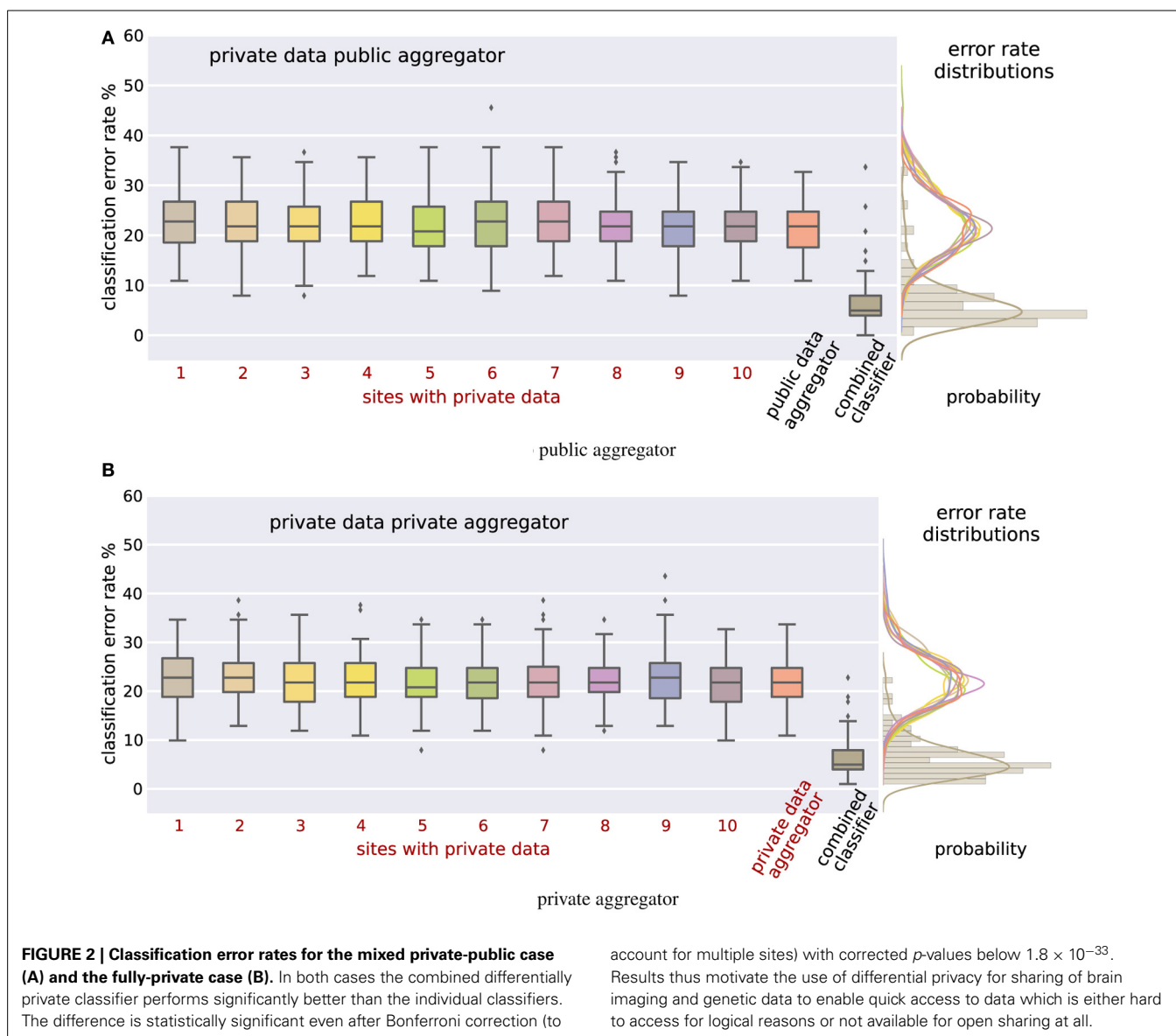
From the perspective of differential privacy it is important to note that the only information that each site releases about its data is the separating hyperplane vector  $\vec{w}_i$  and it does so only once. Considering privacy as a resource a site would want to minimize the loss of this resource. For that, a single release of information in our scheme is better than multiple exchanges in any of the

iterative approaches (e.g., Gabay and Mercier, 1976; Zhang et al., 2012a).

We implemented the above system on a neuroimaging dataset (structural MRI scans) with  $N = 10$  private sites. We combined data from four separate schizophrenia studies conducted at Johns Hopkins University (JHU), the Maryland Psychiatric Research Center (MPRC), the Institute of Psychiatry, London, UK (IOP), and the Western Psychiatric Institute and Clinic at the University of Pittsburgh (WPIC) (see Meda et al., 2008). The sample comprised 198 schizophrenia patients and 191 matched healthy controls (Meda et al., 2008). Our implementation relies on the differentially private SVM and logistic regression as described by Chaudhuri et al. (2011) and implementation available online<sup>1</sup>. The differentially private Hubertized SVM in our implementation used regularization parameter  $\lambda = 0.01$ , privacy parameter  $\epsilon = 10$ , and the Huber constant  $h = 0.5$ , while parameters for differentially private logistic regression were set to  $\lambda = 0.01$  and  $\epsilon = 10$  (for details see Chaudhuri et al., 2011). The quality of classification depends heavily on the quality of features; because distributed and differentially private feature learning algorithms are still under development, for the purposes of this example we assume features are given. To learn the features for this demonstration we used a restricted Boltzmann machine (RBM) (Hinton, 2000) with 50 sigmoidal hidden units. For training we have employed an implementation from Nitish Srivastava<sup>2</sup>. We have used  $L_1$ -regularization of the feature matrix  $W(\lambda \|W\|_1)$  ( $\lambda = 0.1$ ) and 50% dropout to encourage sparse features and effectively handle segmented gray matter images of 60465 voxels each. The learning rate parameter was set to 0.01. The weights were updated using the truncated Gibbs sampling method called contrastive divergence (CD) with a single sampling step (CD-1). Further information on RBM model can be found in Hinton (2000) and Hinton et al. (2006). After the RBM was trained we activated all 50 hidden units on each subject's MRI producing a 50 dimensional dataset. Note, no manual feature

<sup>1</sup><http://cseweb.ucsd.edu/~kamalika/code/dperm/>

<sup>2</sup><https://github.com/nitishsrivastava/deepnet>



selection was involved as each and every feature was used. Using these features we repeated the following procedure 100 times:

1. Split the complete set of 389 subjects into class-balanced training and test sets comprising 70% (272 subjects) and 30% (117 subjects) of the data, respectively. The training set was split into  $N + 1 = 11$  class-balanced subsets (sites) of 24 or 25 subjects each.
2. Train a differentially private SVM on  $N = 10$  of these subsets independently (sites with private data).
3. Transform the data of the 11th subset (aggregator) using the trained SVM classifiers (as described above).
4. Train both a differentially private classifier (fully-private) and a standard logistic regression classifier (public-use) on the transformed dataset (combined classifier).
5. Compute the individual error rates on the test set for each of the  $N = 10$  sites. Compute the error rates of a (differentially

private) SVM trained on the data of 11th dataset and the aggregate classifier in Equation (8) that uses differentially private results from all of the sites.

The results that we obtained in this procedure are summarized in **Figure 2** for the mixed private-public (**Figure 2A**) as well as the fully-private (**Figure 2B**) cases. The 10 sites with private data all have base-line classification error rates of a little over 20%, indicating the relative difficulty of this classification task and highlighting the effect of the noise added for differential privacy. That is, on their own, each site would only be able to learn with that level of accuracy. The distribution of the error rates across experiments is given to the right. The last column of each figure shows the error rate of the combined classifier; **Figure 2A** shows the results for a public aggregator, and **Figure 2B** for the private aggregator. In both cases the error rate of the aggregated classifier is around 5%, which is a significant improvement over

a single site. Additionally, the distribution of the error of the combined classifier is more tightly concentrated about its mean. To quantify the significance of the improvement we performed 2-sample *t*-tests for the distribution of the error rates of the combined classifier against error rate distributions of classifiers produced at individual sites. The largest Bonferroni corrected *p*-value was  $1.8 \times 10^{-33}$ . The experiments clearly show the benefits of sharing the results of differentially private computations over simply using the data at a single site. Even though the classifier that each site shares is a noisy version of what they could learn privately and thus less accurate, aggregating noisy classifiers produces at multiple sites dramatically lowers the resulting error.

#### 4. DISCUSSION

Data sharing interfaces must take into account the realities of neuroimaging studies—current efforts have been very focused on the data structures and ability to query, retrieve and share complex and multi-modal datasets, usually under a fixed model of centralized warehousing, archiving, and privacy restrictions. There has been a remarkable lack of focus on the very important issues surrounding the lack of DUAs in older studies and also the privacy challenges which are growing as more data becomes available and predictive machine learning becomes more common.

We must consider several interlocking aspects when choosing a data sharing framework and the technology to enable it. Neuroimaging and genetics data present significant unique challenges for privacy. Firstly, this kind of data is very different from that considered by many works on privacy—images and sequence data are very high-dimensional and highly identifiable, which may set limits on what we expect to be achievable. Secondly, we must determine the data sharing structure—how is data being shared, and to whom. Institutional data holders may allow other institutions, individual researchers, or the public to access their data. The structure of the arrangement can inform which privacy technology is appropriate (Jiang et al., 2013). Thirdly, almost all privacy-preserving data sharing and data mining technologies are still under active research development and are not at the level of commercially deployed security technologies such as encryption for e-Commerce. A privacy-preserving computation model should be coupled with a legal and policy framework that allows enforcement in the case of privacy breaches. In our proposed model, sites can participate in a consortium in which only differentially private data derivatives are shared. By sharing access to the data, rather than the data itself, we mitigate the current proliferation of individually-generated DUAs, by allowing local data holders to maintain more control.

There are a number of challenges in building robust and scalable data sharing systems for neuroinformatics. On the policy side, standards and best practices should be established for data sharing within and across research consortia. For example, one major challenge is attribution and proper crediting for data used in large-scale studies. On the technology side, building federated data sharing systems requires additional fault-tolerance, security, and more sophisticated role-management than is typically found in the research environment. As noted by Haeberlen et al. (2011) implementing a differentially private system introduces

additional security challenges without stricter access controls. Assigning different trust levels for different users (Vinterbo et al., 2012), managing privacy budgets, and other data governance policy issues can become quite complicated with differential privacy. On the statistical side, we must extend techniques from meta-analyses to interpret statistics computed from data sampled under heterogeneous protocols. However, we believe these challenges can be overcome so that researchers can more effectively collaborate and learn from larger populations.

#### FUNDING

This work was supported by the National Institutes of Health via awards NIMH U01MH097435 (Lei Wang, PI) to Jessica A. Turner and NIBIB R01EB005846 and COBRE 5P20RR021938/P20GM103472 to Vince D. Calhoun.

#### ACKNOWLEDGMENTS

Anand D. Sarwate would like to acknowledge helpful discussions with K. Chaudhuri. All of the authors would like to thank the reviewers, whose extensive feedback helped to significantly improve the manuscript.

#### REFERENCES

- Allen, E. A., Erhardt, E. B., Damaraju, E., Gruner, W., Segall, J. M., Silva, R. F., et al. (2011). A baseline for the multivariate comparison of resting state networks. *Front. Syst. Neurosci.* 5:2. doi: 10.3389/fnsys.2011.00002
- Arbabshirani, M. R., Kiehl, K., Pearlson, G., and Calhoun, V. D. (2013). Classification of schizophrenia patients based on resting-state functional network connectivity. *Front. Neurosci.* 7:133. doi: 10.3389/fnins.2013.00133
- Bießmann, F., Plis, S., Meinecke, F. C., Eichele, T., and Müller, K. R. (2011). Analysis of multimodal neuroimaging data. *IEEE Rev. Biomed. Eng.* 4, 6. doi: 10.1109/RBME.2011.2170675
- Bridwell, D. A., Wu, L., Eichele, T., and Calhoun, V. D. (2013). The spatio-spectral characterization of brain networks: fusing concurrent EEG spectra and fMRI maps. *Neuroimage* 69, 101–111. doi: 10.1016/j.neuroimage.2012.12.024
- Chaudhuri, K., and Mishra, N. (2006). “When random sampling preserves privacy,” in *Advances in Cryptology - CRYPTO 2006*. Lecture notes in computer science, Vol. 4117, ed C. Dwork (Berlin: Springer-Verlag), 198–213. doi: 10.1007/11818175\_12
- Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *J. Mach. Learn. Res.* 12, 1069–1109.
- Chaudhuri, K., Sarwate, A. D., and Sinha, K. (2013). A near-optimal algorithm for differentially-private principal components. *J. Mach. Learn. Res.* 14, 2905–2943.
- Chen, J., Calhoun, V. D., Pearlson, G. D., Perrone-Bizzozero, N., Sui, J., Turner, J. A., et al. (2013). Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference. *Neuroimage* 83, 384–396. doi: 10.1016/j.neuroimage.2013.05.073
- Couzin, J. (2008). Genetic privacy. Whole-genome data not anonymous, challenging assumptions. *Science* 321, 1728. doi: 10.1126/science.321.5894.1278
- Deshpande, G., Libero, L., Sreenivasan, K. R., Deshpande, H., and Kana, R. K. (2013). Identification of neural connectivity signatures of autism using machine learning. *Front. Hum. Neurosci.* 7:670. doi: 10.3389/fnhum.2013.00670
- Dinur, I., and Nissim, K. (2003). “Revealing information while preserving privacy,” in *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (New York, NY: ACM), 202–210.
- Duchi, J., Jordan, M., and Wainwright, M. (2012). “Privacy aware learning,” in *Advances in Neural Information Processing Systems 25*, eds P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (La Jolla, CA: Neural Information Processing Systems Foundation), 1439–1447.
- Duchi, J., Wainwright, M. J., and Jordan, M. (2013). Local privacy and minimax bounds: sharp rates for probability estimation. *Adv. Neural Inform. Process. Syst.* 26, 1529–1537.

- Dwork, C., and Lei, J. (2009). "Differential privacy and robust statistics," in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC '09)* (New York, NY: ACM), 371–380. doi: 10.1145/1536414.1536466
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*. Lecture notes in computer science, Vol. 3876, eds S. Halevi and T. Rabin (Berlin, Heidelberg: Springer), 265–284.
- Dwork, C., Rothblum, G., and Vadhan, S. (2010). "Boosting and differential privacy," in *51st Annual IEEE Symposium on Foundations of Computer Science (FOCS '10)* (Las Vegas, NV), 51–60.
- Dwork, C., and Smith, A. (2009). Differential privacy for statistics: what we know and what we want to learn. *J. Privacy Confidential*. 1, 135–154.
- Fan, L., and Xiong, L. (2012). "Real-time aggregate monitoring with differential privacy," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)* (New York, NY: ACM), 2169–2173.
- Fennema-Notestine, C., Gamst, A. C., Quinn, B. T., Pacheco, J., Jernigan, T. L., Thal, L., et al. (2007). Feasibility of multi-site clinical structural neuroimaging studies of aging using legacy data. *Neuroinformatics* 5, 235–245. doi: 10.1007/s12021-007-9003-9
- Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: a survey of recent developments. *ACM Comput. Surv.* 42, 14:1–14:53. doi: 10.1201/9781420091502
- Gabay, D., and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* 2, 17–40. doi: 10.1016/0898-1221(76)90003-1
- Ganta, S. R., Kasiviswanathan, S. P., and Smith, A. (2008). "Composition attacks and auxiliary information in data privacy," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)* (New York, NY: ACM), 265–273. doi: 10.1145/1401890.1401926
- Geng, Q., and Viswanath, P. (2012). *The Optimal Mechanism in Differential Privacy*. Technical Report arXiv:1212.1186.
- Geng, Q., and Viswanath, P. (2013). *The Optimal Mechanism in  $(\epsilon, \delta)$ -Differential Privacy*. Technical Report arXiv:1305.1330.
- Ghosh, A., and Roth, A. (2011). "Selling privacy at auction," in *Proceeding of the 12th ACM Conference on Electronic Commerce (EC '11)* (New York, NY: ACM), 199–208.
- Ghosh, A., Roughgarden, T., and Sundararajan, M. (2012). Universally utility-maximizing privacy mechanisms. *SIAM J. Comput.* 41, 1673–1693. doi: 10.1137/09076828X
- Girirajan, S., Dennis, M. Y., Baker, C., Malig, M., Coe, B. P., Campbell, C. D., et al. (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am. J. Hum. Genet.* 92, 221–237. doi: 10.1016/j.ajhg.2012.12.016
- Gupte, M., and Sundararajan, M. (2010). "Universally optimal privacy mechanisms for minimax agents," in *ACM SIGMOD Symposium on Principles of Database Systems (PODS)* (New York, NY: ACM), 135–146.
- Gymrek, M., McGuire, A. L., Folan, D., Halperin, E., and Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science* 339, 321–324. doi: 10.1126/science.1229566
- Haeberlen, A., Pierce, B. C., and Narayan, A. (2011). "Differential privacy under fire," in *Proceedings of the 20th USENIX Conference on Security* (Berkeley, CA: USENIX Association).
- Hardt, M., Ligett, K., and McSherry, F. (2012). "A simple and practical algorithm for differentially private data release," in *Advances in Neural Information Processing Systems*, Vol. 25, eds P. Bartlett, F. Pereira, C. Burges, L. Bottou and K. Weinberger (La Jolla, CA: Neural Information Processing Systems Foundation), 2348–2356.
- Hardt, M., and Roth, A. (2012). "Beating randomized response on incoherent matrices," in *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC '12)* (New York, NY: ACM), 1255–1268.
- Hardt, M., and Roth, A. (2013). "Beyond worst-case analysis in private singular vector computation," in *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC '13)* (New York, NY: ACM), 331–340. doi: 10.1145/2488608.2488650
- Hilbar, D., Calhoun, V., and Enigma Consortium (2013). "ENIGMA2: genome-wide scans of subcortical brain volumes in 16,125 subjects from 28 cohorts worldwide," in *19th Annual Meeting of the Organization for Human Brain Mapping* (Seattle, WA).
- Hinton, G. (2000). Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 2002. doi: 10.1162/089976602760128018
- Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527
- Homer, N., Szlinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet.* 4:e1000167. doi: 10.1371/journal.pgen.1000167
- Huang, Y., Malka, L., Evans, D., and Katz, J. (2011). "Efficient privacy-preserving biometric identification," in *Proceedings of the 18th Network and Distributed System Security Conference (NDSS 2011)*.
- Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27, 685–691. doi: 10.1002/jmri.21049
- Jain, P., Kothari, P., and Thakurta, A. (2012). "Differentially private online learning," in *Proceedings of the 25th Annual Conference on Learning Theory (COLT '12)*. JMLR workshop and conference proceedings, Vol. 23 (Scotland: Edinburgh), 24.1–24.34.
- Jain, P., and Thakurta, A. (2013). "Differentially private learning with kernels," in *Proceedings of The 30th International Conference on Machine Learning (ICML)*. JMLR: workshop and conference proceedings, Vol. 28, eds S. Dasgupta and D. McAllester (Beijing: International Machine Learning Society), 118–126.
- Jain, P., and Thakurta, A. (2014). "(near) dimension independent risk bounds for differentially private learning," in *Proceedings of the 31st International Conference on Machine Learning* (Atlanta, GA).
- Jiang, X., Sarwate, A. D., and Ohno-Machado, L. (2013). Privacy technology to share data for comparative effectiveness research: a systematic review. *Med. Care* 51, S58–S65. doi: 10.1097/MLR.0b013e31829b1d10
- Karwa, V., Raskhodnikova, S., Smith, A., and Yaroslavtsev, G. (2011). Private analysis of graph structure. *Proc. VLDB Endowment* 4, 1146–1157.
- Kasiviswanathan, S., Nissim, K., Raskhodnikova, S., and Smith, A. (2013). "Analyzing graphs with node differential privacy," in *Proceedings of the 10th Theory of Cryptography Conference (TCC)* (Tokyo), 457–476. doi: 10.1007/978-3-642-36594-2\_26
- Kifer, D., and Machanavajjhala, A. (2011). "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data* (New York, NY: ACM), 193–204. doi: 10.1145/1989323.1989345
- Kifer, D., Smith, A., and Thakurta, A. (2012). "Private convex empirical risk minimization and high-dimensional regression," in *Proceedings of the 25th Annual Conference on Learning Theory (COLT '12)*. JMLR Workshop and Conference Proceedings, Vol. 23, eds S. Mannor, N. Srebro and R. C. Williamson (Scotland: Edinburgh), 25.1–25.40.
- Lei, J. (2011). "Differentially private M-estimators," in *Advances in Neural Information Processing Systems* 24, eds J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger (La Jolla, CA: Neural Information Processing Systems Foundation), 361–369.
- Le Ny, J., and Pappas, G. J. (2012a). "Differentially private filtering," in *Proceedings of the 51st Conference on Decision and Control (CDC)* (Maui, HI), 3398–3403.
- Le Ny, J., and Pappas, G. J. (2012b). "Differentially private Kalman filtering," in *Proceedings of the 50th Annual Allerton Conference on Communications, Control and Computing* (Monticello, IL), 1618–1625.
- Lindell, Y., and Pinkas, B. (2009). Secure multiparty computation for privacy-preserving data mining. *J. Priv. Confidential*. 1, 59–98.
- Li, N., Li, T., and Venkatasubramanian, S. (2007). "t-closeness: privacy beyond k-anonymity and  $\ell$ -diversity," in *IEEE 23rd International Conference on Data Engineering (ICDE)* (Istanbul), 106–115.
- Liu, J., Ghassemi, M. M., Michael, A. M., Boutte, D., Wells, W., Perrone-Bizzozero, N., et al. (2012). An ica with reference approach in identification of genetic variation and associated brain networks. *Front. Hum. Neurosci.* 6:21. doi: 10.3389/fnhum.2012.00021
- Lowe, H. J., Ferris, T. A., Hernandez, P. M., and Weber, S. C. (2009). "Stride—an integrated standards-based translational research informatics platform," in *Proceedings of the 2009 AMIA Annual Symposium* (San Francisco, CA), 391–395.

- Ludman, E. J., Fullerton, S. M., Spangler, L., Trinidad, S. B., Fujii, M. M., Jarvik, G. P., et al. (2010). Glad you asked: participants' opinions of re-consent for dbGaP data submission. *J. Empir. Res. Hum. Res. Ethics* 5, 9–16. doi: 10.1525/jer.2010.5.3.9
- Machanavajjhala, A., Kifer, D., Abowd, J. M., Gehrke, J., and Vilhuber, L. (2008). "Privacy: theory meets practice on the map," in *IEEE 24th International Conference on Data Engineering (ICDE)* (Cancun), 277–286.
- Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkatasubramanian, M. (2007). *l*-diversity: privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data* 1, 3. doi: 10.1145/1217299.1217302
- Malin, B. (2008). *k*-unlinkability: a privacy protection model for distributed data. *Data Knowl. Eng.* 64, 294–311. doi: 10.1016/j.datak.2007.06.016
- McGuire, A. L., Basford, M., Dressler, L. G., Fullerton, S. M., Koenig, B. A., Li, R., et al. (2011). Ethical and practical challenges of sharing data from genome-wide association studies: the eMERGE consortium experience. *Genome Res.* 21, 1001–1007. doi: 10.1101/gr.120329.111
- McSherry, F. (2010). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Commun. ACM* 53, 89–97. doi: 10.1145/1810891.1810916
- McSherry, F., and Mahajan, R. (2010). "Differentially-private network trace analysis," in *Proceedings of SIGCOMM* (New Delhi).
- Meda, S. A., Giuliani, N. R., Calhoun, V. D., Jagannathan, K., Schretlen, D. J., Pulver, A., et al. (2008). A large scale ( $n = 400$ ) investigation of gray matter differences in schizophrenia using optimized voxel-based morphometry. *Schizophr. Res.* 101, 95–105. doi: 10.1016/j.schres.2008.02.007
- Mennes, M., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2013). Making data sharing work: the fcp/indi experience. *Neuroimage* 82, 683–691. doi: 10.1016/j.neuroimage.2012.10.064
- Mohan, P., Thakurta, A., Shi, E., Song, D., and Culler, D. (2012). "GUPT: privacy preserving data analysis made easy," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (New York, NY: ACM), 349–360. doi: 10.1145/2213836.2213876
- Murphy, S. N., and Chueh, H. (2002). "A security architecture for query tools used to access large biomedical databases," in *AMIA, Fall Symposium 2002*, 552–556.
- Murphy, S. N., Mendis, M. E., Berkowitz, D. A., Kohane, I., and Chueh, H. C. (2006). "Integration of clinical and genetic data in the i2b2 architecture," in *Proceedings of the 2006 AMIA Annual Symposium* (Washington, DC), 1040.
- Narayanan, A., and Shmatikov, V. (2008). "Robust de-anonymization of large sparse datasets," in *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (Oakland, CA), 111–125. doi: 10.1109/SP.2008.33
- National Research Council. (1997). *The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition*. Washington, DC: The National Academies Press.
- Nikolaenko, V., Weinsberg, U., Ioannidis, S., Joye, M., Boneh, D., and Taft, N. (2013). "Privacy-preserving ridge regression on hundreds of millions of records," in *IEEE Symposium on Security and Privacy* (San Francisco, CA), 334–348.
- Nissim, K., and Brenner, H. (2010). "Impossibility of differentially private universally optimal mechanisms," in *IEEE Symposium on Foundations of Computer Science (FOCS)* (Las Vegas, NV).
- Nissim, K., Raskhodnikova, S., and Smith, A. (2007). "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing (STOC '07)* (New York, NY: ACM), 75–84. doi: 10.1145/1250790.1250803
- Oh, S., and Viswanath, P. (2013). *The composition theorem for differential privacy*. Technical Report arXiv:1311.0776 [cs.DS].
- Plis, S. M., Calhoun, V. D., Eichele, T., Weisend, M. P., and Lane, T. (2010). MEG and fMRI fusion for nonlinear estimation of neural and BOLD signal changes. *Front. Neuroinform.* 4:12. doi: 10.3389/fninf.2010.00114
- Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* 7:12. doi: 10.3389/fninf.2013.00012
- Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., et al. (2012). Data sharing in neuroimaging research. *Front. Neuroinform.* 6:9. doi: 10.3389/fninf.2012.00009
- Potkin, S. G., and Ford, J. M. (2009). Widespread cortical dysfunction in schizophrenia: the FBIRN imaging consortium. *Schizophr. Bull.* 35, 15–18. doi: 10.1093/schbul/sbn159
- Rastogi, V., Hay, M., Miklau, G., and Suciu, D. (2009). "Relationship privacy: output perturbation for queries with joins," in *Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '09)* (New York, NY: ACM), 107–116. doi: 10.1145/1559795.1559812
- Reed, J., and Pierce, B. C. (2010). "Distance makes the types grow stronger: a calculus for differential privacy," in *ACM SIGPLAN International Conference on Functional Programming (ICFP)* (Baltimore, MD).
- Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* 45, 1150–1159. doi: 10.1038/ng.2742
- Roy, I., Setty, S. T. V., Kilzer, A., Shmatikov, V., and Witchel, E. (2010). "Airavat: security and privacy for MapReduce," in *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation (NSDI '10)* (Berkeley, CA: USENIX Association).
- Rubinstein, B. I. P., Bartlett, P. L., Huang, L., and Taft, N. (2012). Learning in a large function space: privacy-preserving mechanisms for SVM learning. *J. Priv. Confident.* 4, 65–100.
- Sadeghi, A.-R., Schneider, T., and Wehrenberg, I. (2010). "Efficient privacy-preserving face recognition," in *Information, Security and Cryptology – ICISC 2009*. Lecture notes in computer science, Vol. 5984, eds D. Lee and S. Hong (Berlin: Springer), 229–244. doi: 10.1007/978-3-642-14423-3\_16
- Sarwate, A. D., and Chaudhuri, K. (2013). Signal processing and machine learning with differential privacy: theory, algorithms, and challenges. *IEEE Signal Process. Mag.* 30, 86–94. doi: 10.1109/MSP.2013.2259911
- Schadt, E. E., Woo, S., and Hao, K. (2012). Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.* 44, 603–608. doi: 10.1038/ng.2248
- Schelenz, P. D., Klases, M., Reese, B., Regenbogen, C., Wolf, D., Kato, Y., et al. (2013). Multisensory integration of dynamic emotional faces and voices: method for simultaneous EEG-fMRI measurements. *Front. Hum. Neurosci.* 7:729. doi: 10.3389/fnhum.2013.00729
- Smith, A. (2011). "Privacy-preserving statistical estimation with optimal convergence rates," in *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC '11)* (New York, NY: ACM), 813–822.
- Smith, A., and Thakurta, A. (2013). "Differentially private feature selection via stability arguments, and the robustness of LASSO," in *Conference on Learning Theory*. JMLR: workshop and conference proceedings, Vol. 30, 1–32.
- Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics* 25, 98–110. doi: 10.1111/j.1748-720X.1997.tb01885.x
- Sweeney, L. (2002). *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzz. Knowl. Based Syst.* 10, 557–570. doi: 10.1142/S0218488502001648
- Thakurta, A. G., and Smith, A. (2013). "(Nearly) optimal algorithms for private online learning in full-information and bandit settings," in *Advances in Neural Information Processing Systems 26*, eds C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, 2733–2741.
- Turner, J. A., and Van Horn, J. D. (2012). Electronic data capture, representation, and applications for neuroimaging. *Front. Neuroinform.* 6:16. doi: 10.3389/fninf.2012.00016
- van Erp, T. G., Guella, I., Vawter, M. P., Turner, J., Brown, G. G., McCarthy, G., et al. (2013). Schizophrenia miR-137 locus risk genotype is associated with dorsolateral prefrontal cortex hyperactivation. *Biol. Psychiatry* 75, 398–405. doi: 10.1016/j.biopsych.2013.06.016
- Vinterbo, S. A., Sarwate, A. D., and Boxwala, A. (2012). Protecting count queries in study design. *J. Am. Med. Inform. Assoc.* 19, 750–757. doi: 10.1136/amiajn-2011-000459
- Wasserman, L., and Zhou, S. (2010). A statistical framework for differential privacy. *J. Am. Stat. Assoc.* 105, 375–389. doi: 10.1198/jasa.2009.tm08651
- Williams, O., and McSherry, F. (2010). "Probabilistic inference and differential privacy," in *Advances in Neural Information Processing Systems 23*, eds J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, 2451–2459.
- Wolfson, M., Wallace, S., Masca, N., Rowe, G., Sheehan, N., Ferretti, V., et al. (2010). DataSHIELD: resolving a conflict in contemporary bioscience - performing a pooled analysis of individual-level data without sharing the data. *Int. J. Epidemiol.* 39, 1372–1382. doi: 10.1093/ije/dyq111
- Xiao, X., and Tao, Y. (2007). "*m*-invariance: towards privacy preserving republication of dynamic datasets," in *Proceedings of the 2007 ACM SIGMOD*

- International Conference on Management of Data* (New York, NY: ACM), 689–700. doi: 10.1145/1247480.1247556
- Zhang, C., Lee, H., and Shin, K. G. (2012a). “Efficient distributed linear classification algorithms via the alternating direction method of multipliers,” in *International Conference on Artificial Intelligence and Statistics* (La Palma), 1398–1406.
- Zhang, J., Zhang, Z., Xiao, X., Yang, Y., and Winslett, M. (2012b). Functional mechanism: regression analysis under differential privacy. *Proc. VLDB Endowment* 5, 1364–1375.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 December 2013; paper pending published: 20 February 2014; accepted: 19 March 2014; published online: 07 April 2014.

Citation: Sarwate AD, Plis SM, Turner JA, Arbabshirani MR and Calhoun VD (2014) Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. *Front. Neuroinform.* 8:35. doi: 10.3389/fninf.2014.00035

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Sarwate, Plis, Turner, Arbabshirani and Calhoun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.