




# Multiple imputation for analysis of incomplete data in distributed health data networks

Changgee Chang <sup>1</sup>, Yi Deng<sup>2</sup>, Xiaoqian Jiang <sup>3</sup> & Qi Long <sup>1</sup>✉

Distributed health data networks (DHDNs) leverage data from multiple sources or sites such as electronic health records (EHRs) from multiple healthcare systems and have drawn increasing interests in recent years, as they do not require sharing of subject-level data and hence lower the hurdles for collaboration between institutions considerably. However, DHDNs face a number of challenges in data analysis, particularly in the presence of missing data. The current state-of-the-art methods for handling incomplete data require pooling data into a central repository before analysis, which is not feasible in DHDNs. In this paper, we address the missing data problem in distributed environments such as DHDNs that has not been investigated previously. We develop communication-efficient distributed multiple imputation methods for incomplete data that are horizontally partitioned. Since subject-level data are not shared or transferred outside of each site in the proposed methods, they enhance protection of patient privacy and have the potential to strengthen public trust in analysis of sensitive health data. We investigate, through extensive simulation studies, the performance of these methods. Our methods are applied to the analysis of an acute stroke dataset collected from multiple hospitals, mimicking a DHDN where health data are horizontally partitioned across hospitals and subject-level data cannot be shared or sent to a central data repository.

<sup>1</sup>University of Pennsylvania, Philadelphia, PA, USA. <sup>2</sup>Emory University, Atlanta, GA, USA. <sup>3</sup>University of Texas Health Science Center at Houston, Houston, TX, USA. ✉email: [qlong@upenn.edu](mailto:qlong@upenn.edu)

In the past two decades, enormous amounts of health data have been collected and digitized, partly due to increasingly broader adoption of electronic health records (EHRs) by many healthcare systems. Pooling such big health data from multiple institutions such as healthcare systems and health insurance companies into a single database increases sample sizes for subsequent data analyses and, more importantly, the pooled data can provide a more representative sample of a larger population of interest. As such, it offers great promises in improving the validity, robustness and generalizability of research findings. However, pooling data from multiple institutions may not always be feasible or desirable. First, when the amount of data is massive and continues to grow, it may not be feasible or efficient to transmit data between institutions or store all data in one central repository. Second, for big health data, it may be desirable to store them in a distributed fashion and take advantage of advances in parallel computing. Third, most importantly, due to government regulations, institutional policies, and privacy concerns, it may not be possible to transfer big health data at the patient level from one institution to another or there are extremely high hurdles for such data transferring that may take years to clear. For example, Veteran's Health Administration policies require its EHR data to remain only within VA's facilities. In addition, improper disclosure of individual-level data has serious implications, such as discrimination for employment, insurance, or education<sup>1</sup>. In addition, the current standard practice of data de-identification through removing individual identifiers is inadequate for privacy protection in the era of big data, as a large body of research has demonstrated that given some background information of an individual, an adversary can learn (from "de-identified" data) sensitive information about the victim<sup>2–6</sup>.

To address these challenges, distributed health data networks (DHDNs) that can store and analyze EHRs data from multiple sites without sharing individual-level data have drawn increasing interests in recent years<sup>7,8</sup>. Examples of DHDNs<sup>9</sup>, include the vaccine safety datalink, the health care systems research network, the sentinel initiative, and most recently the patient-centered SCALable national network for effectiveness research (pSCANNER)<sup>10</sup> that is part of PCORnet. To enhance scalability and privacy protection in distributed analysis, the PopMedNet platform<sup>11,12</sup> has been developed to provide software enabled governance over shared data. DHDNs eliminate the need to create, maintain, and secure access to central data repositories, minimize the need to disclose protected health information outside the data-owning entity, and mitigate many security, proprietary, legal, and privacy concerns. In this work, we focus on horizontally partitioned data<sup>13</sup>, meaning that different data custodians such as hospitals and healthcare providers have the same set of features for different sets of patients. For example, several healthcare systems are interested in analyzing pooled data from their EHRs to improve the precision and generalizability of analysis results. However, due to the aforementioned concerns, they are not allowed or are reluctant to share individual-level data with others, despite the substantial benefits from such collaboration. DHDNs would lower the hurdles for them to collaborate in a distributed analysis environment<sup>14</sup>, highlighted needed methods contributions to analysis of distributed EHRs data.

As EHRs are collected as part of healthcare delivery, missing data are pervasive in EHRs and DHDNs<sup>8,15</sup>. Missing data problem reduces the usable sample size and hence analysis power. Improper handling of missing data is known to compromise the validity of analysis and yield biased results, and could subsequently lead to inappropriate healthcare and health policy decisions. To choose the best way forward in handling missing data, the pattern and mechanism of missingness need to be considered<sup>16</sup>. Three main missing data mechanisms are missing

completely at random (MCAR), missing at random (MAR), and missing not at random<sup>17</sup>. Most of the existing methods for handling missing data rely on the assumption of MAR, i.e., missingness only depends on observed data, and which is the focus of our current work as well.

Multiple imputation (MI)<sup>17</sup> is arguably the most popular method for handling missing data largely due to its ease of use. MI methods replace each missing value with samples from its posterior predictive distribution. The predictive imputation model is estimated from the observed data, which have no missing values. The missing values are imputed multiple times in order to account for the the uncertainty of imputation, and then each imputed dataset is used to fit the analysis model parameters  $\theta$ <sup>18</sup> proposed a simple method for combining these analysis results from multiple imputed datasets, which is known as Rubin's rule. In the presence of general missing data patterns, the MI by chained equations (MICE) method is widely adopted and has been shown to achieve superior performance in practice<sup>19,20</sup>.

While there has been a large body of literature on handling missing data, there has been little work on handling distributed incomplete data such as missing data in DHDNs. Of note, while pSCANNER<sup>10</sup> has developed a suit of software tools for privacy-preserving distributed data analysis, it currently has no tools for handling distributed missing data.

To enhance protection of patient privacy, we investigate distributed MI methods for handling missing data that do not require sharing individual level data between sites. Under MAR, one straightforward privacy-preserving MI approach for horizontally partitioned incomplete data would be to conduct MI within each institution/site and then perform the distributed analysis. We call this approach the independent MI (iMI). The iMI has a number of limitations. In particular, it fails to leverage data from other sites, which leads to large variability in imputation and loss of power in subsequent analysis. This becomes more pronounced as the proportion of missing data in individual sites increases. In the extreme case when one variable is missing for all observations in a single site, this variable cannot be imputed in that site using the iMI approach. As a result, the data from this site may not be used in any subsequent analysis where that variable is needed<sup>21</sup> proposed a privacy-preserving lazy decision-tree imputation algorithm for data that are horizontally partitioned between two sources. As their algorithm is designed for only single imputation, it is challenging to conduct proper statistical inference such as hypothesis testing using their singly imputed dataset that underestimates the uncertainty of imputation. In addition, it is not directly applicable to general missing data patterns and the case of more than two sources, and their complex decision tree algorithm may overfit the data and may not be communication efficient.

Since communicating data between sites in distributed learning can be a costly operation, we seek to develop communication-efficient distributed MI approaches. The aforementioned naïve iMI approach is communication-efficient as it involves no communication between sites. We propose two additional communication-efficient approaches, inspired by the inference methods for distributed complete data (CD); the average mixture approach (AVGM) and the communication-efficient surrogate likelihood (CSL) approach. In the AVGM approach<sup>22</sup>, each site finds the local estimate using the data available at the site, and then these estimates are averaged to find the global estimate. The CSL approach<sup>23</sup> uses the curvature information from a central site and the pooled derivative at a point near the true parameter. AVGM is expected to perform better when the samples are evenly distributed across the sites, while CSL is expected to perform well when the central site has the majority of samples. We will use these two approaches to develop distributed MI approaches, avgmMI and cslMI, for univariate missing data patterns. In addition, we develop the another

distributed MI method that uses only the aggregated statistics from each site that are sufficient to obtain the same global estimate as if one had access to data pooled from all sites, and we call this method siMI. siMI can be communication efficient for linear regression models but not for nonlinear regression models for which the fitting algorithm is iterative. Of note, similar to iMI, when one variable is missing for all observations in a single site, this variable cannot be imputed in that site using avgmMI and cslMI and hence the data from this site may not be used in any subsequent analysis where that variable is needed. However, siMI would enable the use of data from these sites, which is one advantage of siMI over the other methods. Using these techniques, we also develop distributed MICE methods for general missing data patterns. However, since the standard MICE algorithm involves fitting of the imputation model multiple times, these direct extensions may not be as communication efficient except for iMICE which requires no communication.

Our work represents the first attempt to develop MI methods that allow proper statistical inference such as hypothesis testing in analysis of horizontally partitioned incomplete data in DHDNs. The remainder of the article is organized as follows. In the section “Results”, we assess the strengths and weaknesses of the proposed distributed imputation methods in simulation studies; we then apply the methods to analysis of an acute stroke dataset collected from EHRs of multiple hospitals, mimicking a DHDN setting. The section “Discussion” provides some concluding remarks. In the section “Methods”, we first briefly review the standard MI and the two distributed analysis methods for CD, AVGM, and CSL, respectively, and then present our communication-efficient distributed MI and MICE methods, respectively.

## Results

**Simulation studies.** We conduct simulation studies to investigate strengths and limitations of the four privacy-preserving distributed MI methods described in the section “Methods” under the MAR assumption. We consider a linear regression model as the “analysis model”

$$y = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p + \epsilon, \tag{1}$$

where  $y$  is the  $N \times 1$  vector of responses  $Y$ ,  $x_1, \dots, x_p$  are the  $N \times 1$  covariate vectors for variables  $X_1$  through  $X_p$ ,  $\theta = (\theta_0, \theta_1, \dots, \theta_p)$  denotes the model parameters of interest, and  $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$  is the  $N \times 1$  vector of errors. We investigate both univariate and general missing data patterns. We apply each distributed MI method to the simulated missing data and then fit the analysis model using the imputed data to evaluate the imputation performance in terms of bias and SD of regression coefficient estimates, and communication costs. To benchmark the performance of the distributed MI methods, we compare their results with the results from the CD analysis which fits the analysis model using the full data before missing values are generated, and the results from the complete case analysis which fits the analysis model using only the set of complete cases that have all variables observed after missing values are generated.

In the first scenario, we have two continuous variables ( $p = 2$ ). The first variable  $X_1$  has missing values while  $X_2$  is fully observed. For each subject,  $X_2$  is first generated from a uniform distribution  $\mathcal{U}(-3, 3)$ . Given  $X_2$ , variable  $X_1$  is sampled from a normal distribution with mean  $\mu_X = 0.2 - 0.5X_2$  and variance  $\sigma_X^2 = 1$ . The outcome  $Y$  is generated from  $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, 1)$  and all  $\theta_j = 1 (j = 0, 1, 2)$ . Variable  $X_1$  is missing with probability  $\{1 + \exp(-0.3 + 0.2Y - 0.1X_2)\}^{-1}$ , resulting in approximately 50% of missing rate.

In the second scenario, we make only one change from the first scenario, which is  $X_1$  is now a binary variable. Given  $X_2$ , instead of sampling  $X_1$  from a normal distribution, we generate  $X_1$

from a Bernoulli distribution  $\mathcal{B}(1, p)$  with probability  $p = \{1 + \exp(-0.2 + 0.5X_2)\}^{-1}$ . The outcome variable  $Y$  and the missingness of  $X_1$  are generated in the same way as in the first scenario. The resulting missing rate is about 50%.

The third scenario considers general missing data patterns. We have  $p = 5$  predictor variables, and  $X_1 - X_3$  have missing values. The fully observed variables  $X_4$  and  $X_5$  are independent and identically distributed as  $\mathcal{N}(0, 1)$ . Given  $X_4$  and  $X_5$ , we sample  $X_1 - X_3$  from a multivariate normal distribution  $\mathcal{N}(\mu_X, \Sigma_X)$  with

$$\mu_X = (0.3 - 0.3X_4 - 0.1X_5)\mathbf{1}, \quad \Sigma_X = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The outcome  $Y$  is generated by  $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 + \theta_5 X_5 + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, 1)$  and all  $\theta_j = 1 (j = 0, 1, \dots, 5)$ . Missing values in  $X_1 - X_3$  are generated based on the logistic regression models for the missing indicators  $\delta_1 - \delta_3$ .

$$\text{logit}(\text{Pr}(\delta_1 = 1)) = -1.0 - 0.4Y - 0.1X_4 - 0.2X_5,$$

$$\text{logit}(\text{Pr}(\delta_2 = 1)) = -0.8 - 0.6Y + 0.2X_4 + 0.4X_5,$$

$$\text{logit}(\text{Pr}(\delta_3 = 1)) = -0.8 - 1.0Y + 0.4X_4 + 0.3X_5,$$

resulting in 20% of missing rates for each missing variable and 50% of complete case rate.

Let  $K$  be the number of data sites distributed over the network. We consider two different numbers of sites ( $K = 5, 10$ ) and three different sample sizes  $N = 250, 500, \text{ and } 1000$ . We also look at two different types of distributions among the samples over the sites. In the first type (U), the samples are unevenly distributed. The first site has the majority of the samples and each site except the first has 15 samples only. In the second type (E), the samples are evenly distributed over the  $K$  sites. Table 1 lists all 15 settings of  $K, N$ , and the sample distribution type which are considered in this study. To evaluate the performance, we compute bias, standard deviation (SD), and root mean squared error of the estimates for  $\theta$  from 1000 Monte Carlo datasets, which are defined as  $\text{Bias}(\theta) = \|\mathbb{E}\theta - \theta_0\|_2$ ,  $\text{SD}(\theta) = \sqrt{\mathbb{E}\|\theta - \mathbb{E}\theta\|_2^2}$ ,

and  $\text{rMSE}(\theta) = \sqrt{\mathbb{E}\|\theta - \theta_0\|_2^2}$ , where  $\theta_0$  is the true value of  $\theta$ .

Tables 2–4 summarize the results for scenarios 1–3, respectively. Note that the CC method is biased regardless of the sample size  $N$ , which is as expected since the missing mechanism is not MCAR. Overall, the biases of all MI methods deteriorate as  $N$  decreases and  $K$  increases. However, the changes vary with the method, the type of sample distribution, and the type of imputed variable.

We can see that iMI and iMICE are less biased when the samples are evenly distributed, as each individual imputation model can be fitted stably. In contrast, when most of sites do not have enough samples, the individual and hence the aggregated estimates are less stable. Note that avgmMI and avgmMICE are hardly affected by the type of sample distribution when the missing variable is continuous (scenarios 1 and 3). However, they are substantially influenced when the missing variable is binary (scenario 2). The difference is even bigger when  $K = 10$ . Conversely, cslMI and cslMICE are worse when the samples are evenly distributed, obviously because the sample size at the central site is smaller. Note that they utilize the curvature information from the central site only and the initial estimate is obtained from the central site as well. Therefore, its performance is sensitive to the sample size at the central site. In particular, note that a few cases of cslMICE failed to converge in evenly distributed case when  $K = 10$  and  $N = 250$ . However, the performance of the CSL based methods is comparable to that of siMI and siMICE when the central site has the majority of the samples. Note that the estimates of siMI and siMICE are as

**Table 1 Fifteen different distributions of samples.**

Type	K	N	n <sup>(1)</sup>	n <sup>(2)</sup>	n <sup>(3)</sup>	n <sup>(4)</sup>	n <sup>(5)</sup>	n <sup>(6)</sup>	n <sup>(7)</sup>	n <sup>(8)</sup>	n <sup>(9)</sup>	n <sup>(10)</sup>
-	1	250	250									
-	1	500	500									
-	1	1000	1000									
U	5	250	190	15	15	15	15					
U	5	500	440	15	15	15	15					
U	5	1000	940	15	15	15	15					
U	10	250	115	15	15	15	15	15	15	15	15	15
U	10	500	365	15	15	15	15	15	15	15	15	15
U	10	1000	865	15	15	15	15	15	15	15	15	15
E	5	250	50	50	50	50	50					
E	5	500	100	100	100	100	100					
E	5	1000	200	200	200	200	200					
E	10	250	25	25	25	25	25	25	25	25	25	25
E	10	500	50	50	50	50	50	50	50	50	50	50
E	10	1000	100	100	100	100	100	100	100	100	100	100

Type indicates whether the samples are unevenly (U) distributed or evenly (E) distributed. K is the number of sites. N is the total number of samples.

unbiased as comparable to CD in all settings, although they suffer a bit larger SDs.

Note that iMI and iMICE do not require any communication for imputation. The avgmMI approach only requires two one-way communications; one to fit the imputation model by AVGM and another to deliver the aggregated estimates to all sites for imputation. The cslMI approach requires one more one-way communication; two to fit the imputation model by CSL and another to deliver the estimated estimates to all sites for imputation. However, the cslMI method transmits vectors only in the first two communications, while avgmMI sends an estimate vector and a covariance matrix in every communication. The siMI approach requires as many communications as avgmMI does when the imputation model is a linear regression. However, siMI requires more communications when the imputation model is nonlinear as shown in Table 3.

As we can see in Table 4, the communication costs of the proposed MICE methods are huge except iMICE. Due to the iterative nature of the MI by chained equation, the number of required communications is proportional to the number of imputations, *M*.

**Analysis of real data.** The Georgia Coverdell Acute Stroke Registry (GCASR), covering nearly 80% of acute stroke admissions in the state of Georgia in USA, was set up to monitor and improve the care of acute stroke patients in the prehospital and hospital settings. The GCASR dataset analyzed in this section includes 68,287 patients from 75 hospitals in Georgia with clinically diagnosed acute stroke between 2005 and 2013. The data collected from EHRs in each hospital include a total of 203 variables, many of which have missing values due to various reasons. The goal of our analysis is to fit a linear regression model for assessing the effect of 14 features on the outcome variable of arrival-to-computed tomography time, an important quality indicator for acute stroke care, in the presence of missing data. The features of interest include patient-related characteristics such as age and gender, and pre-hospital-related characteristics such as EMS notification.

To assess the performance of the distributed imputation methods, we consider the case where EHRs data from individual hospitals cannot be pooled or sent to a central data repository, mimicking a DHDN, and we seek to impute missing values in this distributed set-up while protecting data privacy. Among these features of interest, only gender and race are observed for all patients, and the missing rates for the other variables range from 0.04 to 50.73%. Of note, in some hospitals one or more variables

(e.g., NIH stroke score, EMS prenotification, and NPO) are missing for all observations. Since iMICE cannot be used to impute missing values in a hospital with one or more variables missing for all observations, such hospitals were removed in the first set of analyses resulting in 67,944 observations from 66 hospitals. The sample size in each hospital ranges from 18 to 4,333 with median 578. The number of complete cases across all hospitals is 13,353.

As with the simulation study, we used the CC, iMICE, avgmMICE, cslMICE, and siMICE methods. The CD analysis is not applicable to real data analysis. In addition, since the cslMICE approach is sensitive to the sample size of the central site, we consider two versions of cslMICE, namely, cslMICE(*M*) and cslMICE(*m*). For cslMICE(*M*), the central site is chosen to be the one with the most samples (4333). For cslMICE(*m*), the central site is the hospital with the median sample size (578). For each imputation method, we generate *M* = 20 imputed datasets. To benchmark the performance of the distributed MI methods, we also include the results from the complete case analysis, noting that the CD analysis is not applicable in the real data example.

To compare the performance of distributed imputation methods without being complicated by the choice of distributed method for fitting the analysis model, we chose to fit the analysis model using the imputed data pooled across all hospitals. The analysis results from the *M* = 20 imputed datasets are combined using the Rubin’s rule. Specifically, let  $\hat{\theta}_m$  and  $\widehat{\text{Var}}(\hat{\theta}_m)$  be the regression coefficient estimate and its estimated variance (or variance-covariance) from the *m*-th imputed dataset. Then the overall coefficient estimate is given by  $\hat{\theta} = \frac{1}{M} \sum_m \hat{\theta}_m$ , and its estimated variance is given by  $\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{M} \sum_m \widehat{\text{Var}}(\hat{\theta}_m) + \frac{1}{M-1} \sum_m (\hat{\theta}_m - \hat{\theta})(\hat{\theta}_m - \hat{\theta})^T$ .

Figure 1 presents the parameter estimates and associated 95% confidence intervals for each regression coefficient in the linear regression model of interest. Of note, the hospitals in which at least one variable is missing for all observations are removed for iMICE since the missing values in such hospitals cannot be imputed using iMICE. For each method other than siMICE, we counted the number of discrepancies in statistical significance defined at  $\alpha = 0.05$  or in sign/direction of estimated effect compared to siMICE. Table 5 reports the number of discrepancies along with the number of communications required for each imputation method.

Since the results from the siMICE method are the same as the results from the standard MICE using pooled data, the latter is omitted from Fig. 1, and the results from the siMICE method are

**Table 2 Simulation results for scenario 1 where a continuous variable  $X_1$  has missing values.**

Type	K	Method	N = 250				N = 500				N = 1000			
			Bias	SD	rMSE	Com	Bias	SD	rMSE	Com	Bias	SD	rMSE	Com
-	1	CD	0.001	0.103	0.103	0.0	0.001	0.074	0.074	0.0	0.001	0.052	0.052	0.0
U	5	CC	0.104	0.153	0.185	0.0	0.105	0.107	0.150	0.0	0.106	0.075	0.130	0.0
		iMI	0.169	0.224	0.281	0.0	0.105	0.178	0.206	0.0	0.062	0.132	0.146	0.0
		avgmMI	0.022	0.136	0.138	2.0	0.012	0.093	0.094	2.0	0.004	0.065	0.065	2.0
		csiMI	0.002	0.133	0.133	3.0	0.001	0.093	0.093	3.0	0.003	0.065	0.065	3.0
E	10	siMI	0.002	0.131	0.131	2.0	0.002	0.093	0.093	2.0	0.003	0.064	0.064	2.0
		iMI	0.325	0.256	0.414	0.0	0.208	0.215	0.299	0.0	0.133	0.177	0.221	0.0
		avgmMI	0.054	0.140	0.150	2.0	0.029	0.094	0.094	2.0	0.013	0.065	0.067	2.0
		csiMI	0.003	0.138	0.138	3.0	0.002	0.093	0.093	3.0	0.003	0.065	0.065	3.0
E	5	siMI	0.003	0.130	0.130	2.0	0.002	0.092	0.092	2.0	0.004	0.065	0.065	2.0
		iMI	0.068	0.131	0.147	0.0	0.033	0.092	0.098	0.0	0.018	0.064	0.067	0.0
		avgmMI	0.026	0.133	0.135	2.0	0.011	0.092	0.093	2.0	0.004	0.065	0.065	2.0
		csiMI	0.021	0.170	0.171	3.0	0.003	0.111	0.111	3.0	0.002	0.076	0.076	3.0
E	10	siMI	0.004	0.130	0.130	2.0	0.003	0.091	0.091	2.0	0.004	0.065	0.065	2.0
		iMI	0.180	0.140	0.228	0.0	0.076	0.092	0.119	0.0	0.038	0.065	0.075	0.0
		avgmMI	0.063	0.136	0.150	2.0	0.031	0.093	0.098	2.0	0.013	0.065	0.066	2.0
		csiMI	0.150	0.348	0.378	3.0	0.024	0.154	0.156	3.0	0.003	0.091	0.091	3.0
E	10	siMI	0.004	0.130	0.130	2.0	0.003	0.092	0.092	2.0	0.004	0.064	0.065	2.0

Reported are Bias, average bias; SD, Monte Carlo standard deviation; rMSE, root mean squared error; Com, number of communications. N is the total number of samples. See Table 1 for local sample sizes. Results are based on 1000 Monte Carlo datasets.

**Table 3 Simulation results for scenario 2 where a binary variable  $X_1$  has missing values.**

Type	K	Method	N = 250				N = 500				N = 1000			
			Bias	SD	rMSE	Com	Bias	SD	rMSE	Com	Bias	SD	rMSE	Com
-	1	CD	0.013	0.348	0.349	0.0	0.004	0.239	0.239	0.0	0.004	0.166	0.166	0.0
U	5	CC	0.402	0.491	0.635	0.0	0.408	0.347	0.535	0.0	0.407	0.239	0.472	0.0
		iMI	0.077	0.444	0.451	0.0	0.046	0.324	0.327	0.0	0.023	0.227	0.229	0.0
		avgmMI	0.052	0.585	0.587	2.0	0.059	0.441	0.445	2.0	0.056	0.293	0.298	2.0
		csiMI	0.014	0.472	0.472	3.0	0.005	0.332	0.332	3.0	0.004	0.226	0.226	3.0
E	10	siMI	0.014	0.468	0.468	10.7	0.005	0.331	0.331	10.2	0.003	0.232	0.232	9.7
		iMI	0.180	0.404	0.442	0.0	0.105	0.309	0.326	0.0	0.051	0.222	0.228	0.0
		avgmMI	0.115	0.624	0.635	2.0	0.128	0.502	0.518	2.0	0.119	0.333	0.354	2.0
		csiMI	0.014	0.470	0.470	3.0	0.005	0.331	0.331	3.0	0.003	0.230	0.231	3.0
E	5	siMI	0.014	0.469	0.469	10.7	0.005	0.329	0.329	10.2	0.003	0.229	0.229	9.7
		iMI	0.056	0.447	0.451	0.0	0.037	0.324	0.327	0.0	0.019	0.230	0.230	0.0
		avgmMI	0.108	0.524	0.535	2.0	0.047	0.348	0.351	2.0	0.022	0.233	0.234	2.0
		csiMI	0.016	0.533	0.533	3.0	0.005	0.344	0.344	3.0	0.004	0.233	0.233	3.0
E	10	siMI	0.014	0.469	0.469	10.7	0.004	0.330	0.330	10.2	0.002	0.229	0.229	9.7
		iMI	0.169	0.413	0.446	0.0	0.081	0.319	0.329	0.0	0.041	0.226	0.230	0.0
		avgmMI	0.226	0.653	0.691	2.0	0.125	0.386	0.406	2.0	0.054	0.241	0.247	2.0
		csiMI	0.030	0.792	0.793	3.0	0.007	0.446	0.446	3.0	0.003	0.259	0.259	3.0
E	10	siMI	0.015	0.467	0.467	10.7	0.004	0.330	0.330	10.2	0.003	0.228	0.228	9.7

Reported are Bias, average bias; SD, Monte Carlo standard deviation; rMSE, root mean squared error; Com, number of communications. N is the total number of samples. See Table 1 for local sample sizes. Results are based on 1000 Monte Carlo datasets.

**Table 4 Simulation results for scenario 3 where three continuous variables  $X_1$ - $X_3$  have missing values.**

Type	K	Method	N = 250				N = 500				N = 1000			
			Bias	SD	rMSE	Com	Bias	SD	rMSE	Com	Bias	SD	rMSE	Com
U	1	CD	0.004	0.179	0.179	0	0.004	0.127	0.127	0	0.004	0.089	0.089	0
		CC	0.363	0.240	0.435	0	0.365	0.167	0.401	0	0.365	0.117	0.383	0
		iMICE	0.067	0.258	0.267	0	0.037	0.176	0.180	0	0.022	0.121	0.123	0
U	5	avgmMICE	0.011	0.218	0.218	1290	0.006	0.151	0.151	1290	0.004	0.106	0.106	1290
		csIMICE	0.005	0.214	0.214	1935	0.004	0.151	0.151	1935	0.004	0.106	0.106	1935
		siMICE	0.004	0.213	0.213	1290	0.004	0.150	0.150	1290	0.004	0.106	0.106	1290
	10	iMICE	0.146	0.286	0.321	0	0.080	0.201	0.216	0	0.047	0.138	0.146	0
		avgmMICE	0.022	0.221	0.222	1290	0.011	0.153	0.154	1290	0.006	0.106	0.106	1290
		csIMICE	0.005	0.215	0.215	1935	0.004	0.150	0.150	1935	0.005	0.106	0.106	1935
E	5	siMICE	0.004	0.213	0.213	1290	0.004	0.150	0.150	1290	0.005	0.105	0.105	1290
		iMICE	0.027	0.217	0.219	0	0.014	0.151	0.152	0	0.009	0.105	0.106	0
		avgmMICE	0.013	0.215	0.215	1290	0.006	0.151	0.151	1290	0.004	0.105	0.105	1290
	10	csIMICE	0.019	0.243	0.244	1935	0.006	0.156	0.156	1935	0.005	0.107	0.108	1935
		siMICE	0.005	0.213	0.213	1290	0.005	0.150	0.150	1290	0.004	0.105	0.105	1290
		iMICE	0.076	0.228	0.240	0	0.032	0.152	0.155	0	0.016	0.106	0.108	0
E	avgmMICE	0.024	0.219	0.221	1290	0.012	0.151	0.152	1290	0.006	0.106	0.106	1290	
	csIMICE	0.005	0.213	0.213	1290	0.004	0.150	0.150	1290	0.006	0.106	0.105	1290	

Reported are Bias, average bias; SD, Monte Carlo standard deviation; rMSE, root mean squared error; Com, number of communications. The csIMICE method failed in a few cases due to instability when N = 250 samples are evenly (E) distributed over K = 10 sites. N is the total number of samples. See Table 1 for local sample sizes. Results are based on 1000 Monte Carlo datasets.

used to benchmark the other methods. As shown in Table 5, siMICE incurs substantially higher communication costs than the other distributed MI methods. Compared to the results from siMICE, the CC analysis yields substantially different parameter estimates for multiple regression coefficients as well as disagreement in statistical significance or in direction of estimated effect for eight features. This demonstrates the need for adequate handling missing data in the analysis of the GCASR data.

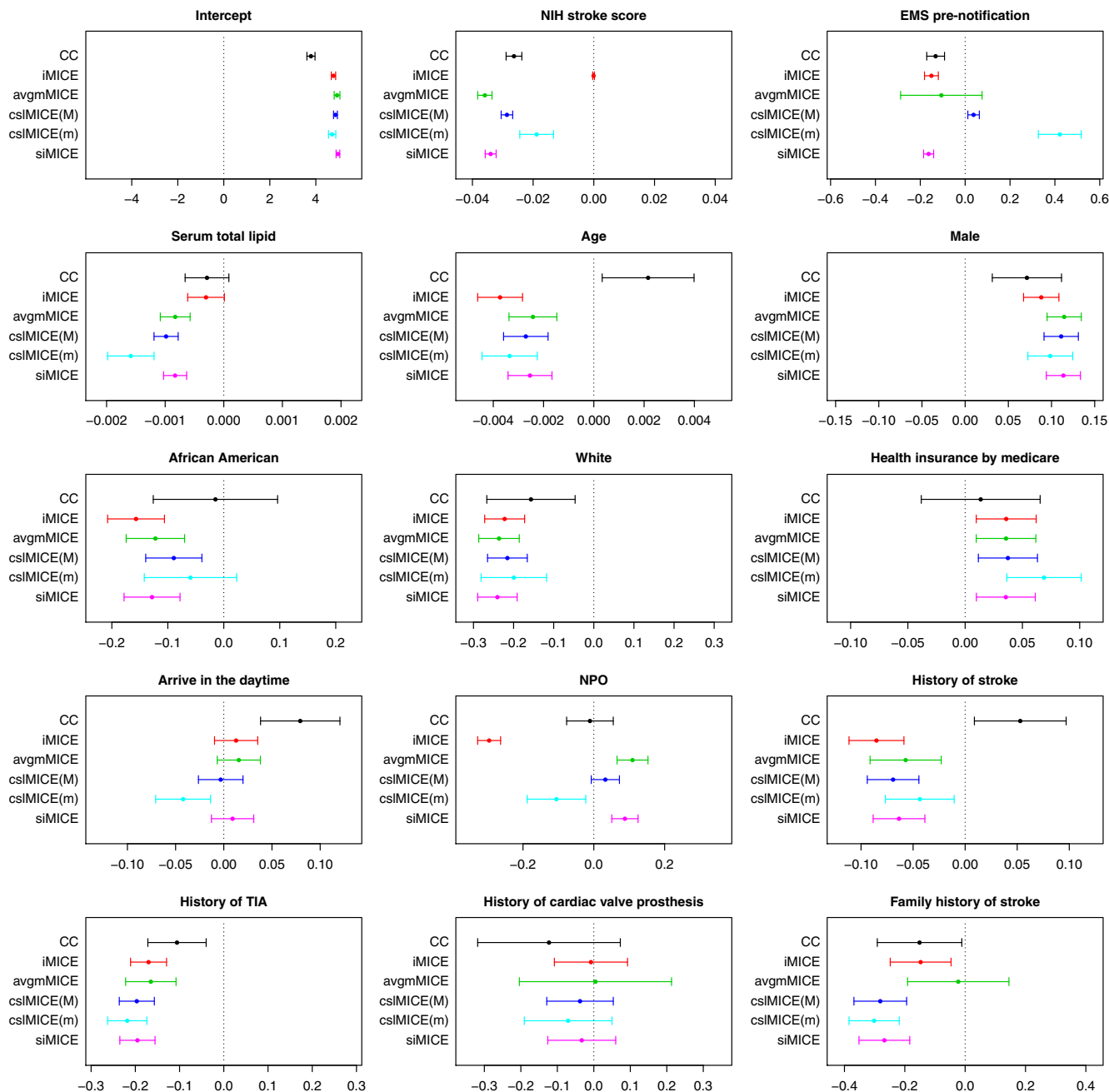
The results from the other distributed MI methods are closer to the results from siMICE than the CC analysis. Though, csIMICE (m), which uses the hospital with the medium sample size as the central site, exhibits the second largest number of discrepancies after CC, including disagreement for four features. While offering substantial savings in communication costs and yielding similar parameter estimates for some features, iMICE shows notable discrepancies for “NIH stroke score,” “Serum total lipid,” and “NPO” when compared to siMICE. On the other hand, csIMICE (M), which uses the hospital with the largest sample size as the central site, and avgmMICE yield the smallest number of discrepancies from siMICE, specifically only two discrepancies as shown in Table 5. In addition, Fig. 1 provides more granular information about comparing csIMICE(M) and avgmMICE with siMICE. The locations of 95% confidence intervals obtained by avgmMICE are most similar to those obtained by siMICE, with the only notable discrepancy for “family history of stroke.” The lengths of confidence intervals obtained by csIMICE(M) are most similar to those obtained by siMICE, which could be attributed to the fact that it uses the curvature information of the central site with a large sample size. As a comparison, avgmMICE yields wider intervals for a number of features including “EMS prenotification,” “history of TIA,” “history of cardiac valve prosthesis,” and “family history of stroke.”

To assess the performance of the distributed MI methods when sample sizes in individual hospitals are moderate to small, we conducted another set of analyses after removing the hospitals with more than  $T$  patients where  $T = 500, 300, 100$ . When  $T = 500$ , the total sample size decreases to 5307 patients from 26 hospitals. The number of patients in each hospital ranges from 18 to 462 with a median of 163 and the number of complete cases is only 362. In this set of analyses, we exclude csIMICE(m). As  $T$  decreases and, in other words, the number of patients per hospital continues to decrease, the discrepancies between the results from siMICE and the results from the other distributed MI methods including iMICE, avgmMICE and csIMICE(M) become greater. Particularly, the discrepancies between siMICE and csIMICE(M) tend to grow faster than the discrepancies between siMICE and avgmMICE, suggesting that csIMICE(M) is more sensitive to moderate to small sample sizes in all sites than avgmMICE.

**Discussion**

In this paper, we consider the problem of distributed incomplete data where data from multiple sites are not allowed to be combined, due to institutional policies or privacy concerns. We have developed and investigated four MI approaches that allow proper statistical inference such as hypothesis testing in analysis of horizontally partitioned incomplete data in DHDNs.

Our numerical experiments provide insights into the strengths and weaknesses of these methods. The proposed distributed imputation methods except for iMI/iMICE enable the use of data from all sites including sites with one or more variables missing for all observations. siMI has been shown in our numerical studies to yield comparable performance as the standard MI using pooled data, but it is not communication efficient for generalized linear imputation models. While csIMI and avgmMI are more communication efficient for all imputation models, their



**Fig. 1 Forest plot for analysis results of the GCASR data.** The parameter estimates (dots) and associated 95% confidence intervals (whiskers) for each regression coefficient including the intercept are compared between all the methods. The hospitals in which at least one variable is missing for all observations have been removed for iMICE. The plots are based on 67,944 observations from 66 hospitals. The sample size in each hospital ranges from 18 to 4333 with median 578.

**Table 5 Comparisons of analysis results of the GCASR data.**

	CC	iMICE	avgmMICE	csMICE(M)	csMICE(m)	siMICE
# of communications	0	0	4730	7095	7095	25,397
# of discrepancies	8	3	2	2	4	-

Reported are the communication costs and the number of discrepancies in statistical significance defined at  $\alpha = 0.05$  or in sign/direction of estimated effect compared against siMICE. The hospitals in which at least one variable is missing for all observations are removed for iMICE.

performance may be sensitive to sample sizes in individual sites. In particular, the performance of csIMI may become unstable as the sample size in the central site becomes small to moderate. While avgmMI is less sensitive to small sample sizes, our simulations show that it tends to yield larger bias when imputing binary variables. Of note, siMI may be particularly appealing when analyzing data for uncommon diseases for which the sample size can be small in each individual site and missing data can further complicate data analysis. On the other hand, given that existing networks have 5 to upwards of 70 or more sites and can have millions to billions of records, the avgmMI and csIMI approaches can be very appealing options compared to the iMI or siMI approaches if all the data are used in analysis.

Unlike the other proposed methods, iMI requires no communication between sites but may lead to unstable results as the sample sizes in some sites become small to moderate. As shown in the real data example, when one variable, say  $X_1$ , is missing for all observations in a single site,  $X_1$  in that site cannot be imputed using iMI/iMICE and hence the data from the site may not be used in subsequent analysis involving  $X_1$ . The choice of distributed imputation approaches may also depend on whether data heterogeneity across multiple sites can be adequately adjusted for in imputation models. If we are able to adequately account for the heterogeneity by say including covariates that capture the heterogeneity or random effects for sites in imputation models, the siMI, csIMI, and avgmMI methods that borrow information across sites can enhance the efficiency of imputation and hence the power of subsequent analysis of imputed datasets. However, if that is not the case, then the iMI approach may be preferred.

We have investigated the extensions of the distributed MI methods for general missing patterns through the use of chained equations (MICE). Although these methods are privacy-preserving and yield good performance, they are not communication efficient as demonstrated in the numerical experiments. In cases where communication costs are of critical concern, more communication-efficient imputation methods are needed for handling general missing data as potential future work. Another potential limitation is that siMI, csIMI and avgmMI may not always be privacy-preserving as the summary statistics transmitted between individual sites and a central server may still leak individual-level information<sup>24</sup>. Particularly, the siMI method needs to transfer the entire design matrix between sites, which poses higher risk of leaking individual patients' information. To address this issue, a differential privacy step<sup>25</sup> can be added to further strengthen the privacy-preserving property.

In practice, robust imputation methods such as predictive mean matching (PMM) and random forest (RF) imputation are widely used. It is of future interest to develop distributed versions of generic imputation methods include PMM and RF, which, however, can be very challenging. Such distributed generic imputation methods are expected to require additional communication overhead and more general definition of sufficient statistics. Particularly, the information to be exchanged for a distributed generic imputation method needs to be carefully investigated, while taking into account statistical validity, privacy-preserving property, and communication costs.

**Methods**

**Ethical approval.** This study was reviewed by the Institutional Review Board at the University of Pennsylvania which determined that the study does not meet the criteria for human subject research since it involves only secondary analysis of de-identified data from an existing database and does not involve new data collection.

**Notation.** To fix ideas, suppose that we are interested in fitting the analysis model (1) of outcome  $Y$  on  $p$  features  $X_1, \dots, X_p$ , using a random sample of  $N$  observations. We define  $\mathbf{x}_0 = \mathbf{1}$  for the  $N \times 1$  vector of ones and denote the values for the  $i$ -

th individual by  $\tilde{\mathbf{x}}_i = (x_{i0} = 1, x_{i1}, \dots, x_{ip})^T$  and  $y_i$ . Let  $\mathbf{X} = [\mathbf{1} \ \mathbf{x}_1 \ \dots \ \mathbf{x}_p] = [\tilde{\mathbf{x}}_1 \ \dots \ \tilde{\mathbf{x}}_N]^T$  be the  $N \times (p + 1)$  design matrix.

We consider horizontally partitioned data from  $K$  institutions or sites, all of which have the same set of features recorded for all of their subjects.  $\mathbf{y}$  and  $\mathbf{X}$ , known as the "pooled" outcome vector and the "pooled" design matrix respectively, can be decomposed by sites as follows:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \vdots \\ \mathbf{y}^{(K)} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(K)} \end{pmatrix},$$

where  $\mathbf{y}^{(k)}$  and  $\mathbf{X}^{(k)}$  are the data from the  $k$ th site with  $n^{(k)}$  subjects,  $\mathbf{X}^{(k)}$  is an  $n^{(k)} \times (p + 1)$  matrix, and  $N = \sum_{k=1}^K n^{(k)}$ .

We first consider a univariate missing pattern where only  $X_1$  has missing values and the other variables are fully observed where  $N_c$  denotes the number of complete cases. We then consider general missing data patterns.

**Multiple imputation.** An MI method replaces each missing value multiple times from its predictive distribution based on the observed data, accounting for the uncertainty of imputation. Each of the imputed datasets is analyzed separately as if it were fully observed. The results across all imputed datasets are then combined following Rubin's rule. For example, if  $X_1$  which has missing values is continuous, we can use a Bayesian linear regression model for imputation

$$X_1 = \alpha_0 + \alpha_1 Y + \sum_{j=2}^p \alpha_j X_j + \zeta, \tag{2}$$

where  $\zeta \sim \mathcal{N}(0, \tau^2)$  with priors

$$\pi(\tau^2) \propto \mathcal{IG}(1/2, 1/2), \quad \boldsymbol{\alpha} | \tau^2 \sim \mathcal{N}(\mathbf{0}, \tau^2 \lambda^{-1} \mathbf{I}),$$

where  $\mathcal{IG}$  and  $\mathcal{N}$  refer to the inverse gamma distribution and the multivariate Gaussian distribution, respectively. Let  $\mathbf{Z} = [\mathbf{1}, \mathbf{y}, \mathbf{x}_2, \dots, \mathbf{x}_p]$ , and let  $\mathbf{Z}_c$  be the  $N_c \times (p + 1)$  submatrix of  $\mathbf{Z}$  loaded with the complete cases only. Similarly, let  $\mathbf{x}_{1,c}$  be the subvector of  $\mathbf{x}_1$  with the complete cases. The posterior distribution of  $(\tau^2, \boldsymbol{\alpha})$  is given by

$$\begin{aligned} \tau^2 | \mathbf{Z}_c &\sim \mathcal{IG}((N_c + 1)/2, (\text{SSE} + 1)/2), \\ \boldsymbol{\alpha} | \tau^2, \mathbf{Z}_c &\sim \mathcal{N}((\mathbf{Z}_c^T \mathbf{Z}_c + \boldsymbol{\Lambda})^{-1} \mathbf{Z}_c^T \mathbf{x}_{1,c}, \tau^2 (\mathbf{Z}_c^T \mathbf{Z}_c + \boldsymbol{\Lambda})^{-1}), \end{aligned} \tag{3}$$

where  $\text{SSE} = \mathbf{x}_{1,c}^T \mathbf{x}_{1,c} - \mathbf{x}_{1,c}^T \mathbf{Z}_c (\mathbf{Z}_c^T \mathbf{Z}_c + \boldsymbol{\Lambda})^{-1} \mathbf{Z}_c^T \mathbf{x}_{1,c}$ . The MI method samples  $(\tau^2, \boldsymbol{\alpha})$  from Equation (3), imputes the missing values of  $X_1$  according to Eq. (2) with random errors added, and fits the analysis model (1) using the imputed full data. This procedure is repeated multiple times.

When  $X_1$  that has missing values is binary, we can use a Bayesian logistic regression model for imputation with prior  $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1} \mathbf{I})$ . Let  $\hat{\boldsymbol{\alpha}}$  be the maximum A posteriori estimator. Note that, as  $N_c$  tends to infinity, we have  $\text{Cov}(\hat{\boldsymbol{\alpha}}) = (\mathbf{Z}_c^T \mathbf{W}_c \mathbf{Z}_c + \boldsymbol{\Lambda})^{-1} (1 + O(N_c^{-1}))$ , where  $\mathbf{W}$  is a diagonal matrix with  $w_{ii} = \text{expit}(\hat{\boldsymbol{\alpha}}_i^T \boldsymbol{\alpha}) (1 - \text{expit}(\hat{\boldsymbol{\alpha}}_i^T \boldsymbol{\alpha}))$ ,  $\mathbf{W}_c$  is the sub-diagonal-matrix of  $\mathbf{W}$  for the complete cases, and  $\text{expit}(x) = \text{logit}^{-1}(x) = \frac{1}{1 + e^{-x}}$ . Therefore, the MI method samples  $\boldsymbol{\alpha}$  from  $\mathcal{N}(\hat{\boldsymbol{\alpha}}, (\mathbf{Z}_c^T \mathbf{W}_c \mathbf{Z}_c + \boldsymbol{\Lambda})^{-1})$ , imputes the missing values according to the Bernoulli distribution,  $x_{i1} \sim B(1, \text{expit}(\hat{\boldsymbol{\alpha}}_i^T \boldsymbol{\alpha}))$ , and fits the analysis model using the imputed full data. This procedure is repeated multiple times. A regularization parameter  $\lambda$  can be used to avoid numerical difficulties particularly when the sample size at a site is less than the dimension of the parameters. We choose the value of  $\lambda$  to be as small as possible so that the bias caused by regularization can be negligible.

**Communication-efficient inference for distributed CD.** One straightforward approach to analyze distributed data is to transmit the minimally sufficient information from all sites to the central site such that it would enable reproducing the results from analyzing data pooled from all sites. We call this approach the SI (sufficient information) method which will be extended to the sufficient information MI (siMI) in the next section. In linear regression, for example, we only need  $\mathbf{X}^{(k)T} \mathbf{X}^{(k)}$  and  $\mathbf{X}^{(k)T} \mathbf{y}^{(k)}$  to obtain the same least-square estimates for the regression coefficients as if we had the data pooled from all sites. This can be seen from the following equation:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y}) = \left( \sum_k \mathbf{X}^{(k)T} \mathbf{X}^{(k)} \right)^{-1} \left( \sum_k \mathbf{X}^{(k)T} \mathbf{y}^{(k)} \right). \tag{4}$$

In addition, this approach requires only one single communication between each site and a central server where the computations described in Eq. (4) are conducted. To extend this strategy to the generalized linear models (GLMs), we note that a standard algorithm for fitting GLMs goes through Newton iterations, each of which requires the derivative and the curvature information of the global loglikelihood that involves regression coefficients. For example, in the logistic regression, the parameters are



updated as follows for each iteration.

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + s \left( \sum_k \mathbf{X}^{(k)T} \mathbf{W}^{(k)} \mathbf{X}^{(k)} \right)^{-1} \left( \sum_k \mathbf{X}^{(k)T} (\mathbf{y}^{(k)} - \boldsymbol{\pi}^{(k)}) \right), \quad (5)$$

where  $\pi_i^{(k)} = \text{expit}(\tilde{\mathbf{x}}_i^{(k)T} \boldsymbol{\theta}^{(t)})$ ,  $\mathbf{W}^{(k)} = \text{diag}(\boldsymbol{\pi}^{(k)}) \text{diag}(\mathbf{I} - \boldsymbol{\pi}^{(k)})$ , and  $s$  is the step size. For each iteration, the central site has to transfer  $\boldsymbol{\theta}^{(t)}$  to other sites and all sites have to transfer  $\mathbf{X}^{(k)T} \mathbf{W}^{(k)} \mathbf{X}^{(k)}$  and  $\mathbf{X}^{(k)T} (\mathbf{y}^{(k)} - \boldsymbol{\pi}^{(k)})$  back to the central site. Since the number of required round-trip communications is same as that of Newton iterations, the SI method may not be communication efficient. This approach is privacy preserving in the sense that the subject-level data are not shared outside of each site and hence are protected. Again, this approach generates the same results as the analysis of the pooled data.

We also consider two alternative distributed analysis methods that are communication efficient<sup>22</sup> proposed the average mixture algorithm. Each site estimates the model parameters using the data available at the site only, and then combine the estimates to find the global estimate for the parameters. Let  $\hat{\boldsymbol{\theta}}^{(k)}$  be the estimate from the  $k$ th site. Then, we have

$$\hat{\boldsymbol{\theta}}_{\text{avgm}} = \sum_k w_k \hat{\boldsymbol{\theta}}^{(k)}, \quad (6)$$

where  $w_k \geq 0$  and  $\sum_k w_k = 1$ . Note that the weights  $w_k$  should reflect the sample size of each site. This method is communication efficient as it requires a single one-way communication only. The resulting estimator can achieve the best rate of convergence in asymptotics<sup>22</sup>. However, the local estimates can be volatile, especially when the sample size of the site is small. Therefore, the finite sample characteristic of  $\hat{\boldsymbol{\theta}}_{\text{avgm}}$  can be quite different. We call this approach the AVGM (average mixture) algorithm. Jordan et al.<sup>23</sup> proposed using the curvature information from one central site, say site 1, and the global derivative at a point near the true parameter. The estimator is defined as the minimizer of the CSL, which is defined as

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}_1(\boldsymbol{\theta}) - \langle \nabla \mathcal{L}_1(\bar{\boldsymbol{\theta}}) - \nabla \mathcal{L}(\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} \rangle,$$

where  $\mathcal{L}$  is the gross average loglikelihood (loss function),  $\mathcal{L}_1$  is the local average loglikelihood at the central site (site 1),  $\bar{\boldsymbol{\theta}}$  is a point close to the true  $\boldsymbol{\theta}$ , and  $\langle \cdot, \cdot \rangle$  denotes the inner product. Note that it requires one round-trip communication in order to calculate  $\nabla \mathcal{L}(\bar{\boldsymbol{\theta}})$ . The solution achieves the optimal convergence rate if  $\bar{\boldsymbol{\theta}}$  converges fast enough<sup>23</sup>. However, the finite sample performance may deteriorate if the sample size at the central site is small, as it utilizes the curvature information from the central site only and the initial solution  $\bar{\boldsymbol{\theta}}$  is also obtained from the central site only. We call this approach the CSL method.

Jordan et al.<sup>23</sup> proposes multiple versions of CSL methods including the ones that repeat the whole procedure using the current solution as a new initial coefficient  $\bar{\boldsymbol{\theta}}$ . Since those approaches require more communications, we do not consider all of them and restrict our focus on the most communication-efficient version, the one described above. The two communication-efficient methods AVGM and CSL are also privacy-preserving in the sense that the individual level data are not shared between sites.

**Distributed MI for univariate missing data pattern.** One straightforward way to impute missing values for distributed data is to impute missing data in each site separately using a standard MI method. We call this approach the iMI method. When using iMI, no subject-level data are shared across sites and thus no communication is required. As such, it is communication-efficient and privacy-preserving. However, this approach has a number of limitations as discussed in the ‘‘Introduction’’ section.

### Algorithm 1

Independent MI algorithm

---

```

1 for  $k \leftarrow 1$  to  $K$  do
2   Fit the imputation model at site  $k$  to find  $\hat{\boldsymbol{\alpha}}^{(k)}$  and  $\text{Cov}(\hat{\boldsymbol{\alpha}}^{(k)})$ ;
3   Sample  $\boldsymbol{\alpha}_1^{(k)}, \dots, \boldsymbol{\alpha}_M^{(k)}$  independently from  $\mathcal{N}(\hat{\boldsymbol{\alpha}}^{(k)}, \text{Cov}(\hat{\boldsymbol{\alpha}}^{(k)}))$ ;
4 end
5 for  $m \leftarrow 1$  to  $M$  do
6   for  $k \leftarrow 1$  to  $K$  do Impute the missing data at site  $k$  based on  $\boldsymbol{\alpha}_m^{(k)}$ 
7   Fit the analysis model and obtain  $\hat{\boldsymbol{\theta}}_m$  and  $\text{Cov}(\hat{\boldsymbol{\theta}}_m)$ ;
8 end
9 Combine the results by Rubin’s rule to obtain  $\hat{\boldsymbol{\theta}}$  and  $\text{Cov}(\hat{\boldsymbol{\theta}})$ ;

```

---

An alternative approach is to use the SI method to fit a distributed imputation model, which would be equivalent to fit the imputation model using data pooled from all sites. The imputation approach requires as many communications as the number of communications required by the SI method plus a one-way communication to deliver the sampled imputation model parameters to all sites. Hence, this method is less communication-efficient than the other methods considered, unless the imputation model is a linear regression. However, as it uses the full information, it is expected to yield the best imputation performance. We call this approach the siMI method, which will be used as a benchmark for assessing imputation performance in our numerical studies.

### Algorithm 2

Sufficient information MI algorithm

---

```

1 Fit the global imputation model using the SI method to find  $\hat{\boldsymbol{\alpha}}_{si}$  and  $\text{Cov}(\hat{\boldsymbol{\alpha}}_{si})$ ;
2 Sample  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_M$  independently from  $\mathcal{N}(\hat{\boldsymbol{\alpha}}_{si}, \text{Cov}(\hat{\boldsymbol{\alpha}}_{si}))$ ;
3 Send  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_M$  to all sites;
4 for  $m \leftarrow 1$  to  $M$  do
5   for  $k \leftarrow 1$  to  $K$  do Impute the missing data at site  $k$  based on  $\boldsymbol{\alpha}_m$ ;
6   Fit the analysis model and obtain  $\hat{\boldsymbol{\theta}}_m$  and  $\text{Cov}(\hat{\boldsymbol{\theta}}_m)$ ;
7 end
8 Combine the results by Rubin’s rule to obtain  $\hat{\boldsymbol{\theta}}$  and  $\text{Cov}(\hat{\boldsymbol{\theta}})$ ;

```

---

We also develop two communication-efficient distributed MI methods, namely, avgmMI and csIMI, which adapt the communication-efficient methods described above. The avgmMI method fits the distributed imputation model using the AVGM method, in which the weights are chosen to be proportional to the number of complete cases in each site.

$$\hat{\boldsymbol{\alpha}}_{\text{avgm}} = \frac{1}{N_c} \sum_k n_c^{(k)} \hat{\boldsymbol{\alpha}}^{(k)},$$

where the covariance matrix of  $\hat{\boldsymbol{\alpha}}_{\text{avgm}}$  is given by

$$\text{Cov}(\hat{\boldsymbol{\alpha}}_{\text{avgm}}) = \frac{1}{N_c^2} \sum_k n_c^{(k)2} \text{Cov}(\hat{\boldsymbol{\alpha}}^{(k)}).$$

Since each site needs to transmit two quantities  $\hat{\boldsymbol{\alpha}}^{(k)}$  and  $\text{Cov}(\hat{\boldsymbol{\alpha}}^{(k)})$  to the central site, avgmMI requires two one-way communications. This is a huge advantage over the siMI method in terms of communication cost except when the imputation model is a linear regression.

### Algorithm 3

AVGM MI algorithm

---

```

1 for  $k \leftarrow 1$  to  $K$  do
2   Find the estimates  $\hat{\boldsymbol{\alpha}}^{(k)}$  and  $\text{Cov}(\hat{\boldsymbol{\alpha}}^{(k)})$  at site  $k$ ;
3   Send  $\hat{\boldsymbol{\alpha}}^{(k)}$  and  $\text{Cov}(\hat{\boldsymbol{\alpha}}^{(k)})$  to the central site;
4 end
5  $\hat{\boldsymbol{\alpha}}_{\text{avgm}} \leftarrow \frac{1}{N_c} \sum_k n_c^{(k)} \hat{\boldsymbol{\alpha}}^{(k)}$ ;
6  $\text{Cov}(\hat{\boldsymbol{\alpha}}_{\text{avgm}}) \leftarrow \frac{1}{N_c^2} \sum_k n_c^{(k)2} \text{Cov}(\hat{\boldsymbol{\alpha}}^{(k)})$ ;
7 Sample  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_M$  independently from  $\mathcal{N}(\hat{\boldsymbol{\alpha}}_{\text{avgm}}, \text{Cov}(\hat{\boldsymbol{\alpha}}_{\text{avgm}}))$ ;
8 Send  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_M$  to all sites;
9 for  $m \leftarrow 1$  to  $M$  do
10   for  $k \leftarrow 1$  to  $K$  do Impute the missing data at site  $k$  based on  $\boldsymbol{\alpha}_m$ 
11   Fit the analysis model and obtain  $\hat{\boldsymbol{\theta}}_m$  and  $\text{Cov}(\hat{\boldsymbol{\theta}}_m)$ ;
12 end
13 Combine the results by Rubin’s rule to obtain  $\hat{\boldsymbol{\theta}}$  and  $\text{Cov}(\hat{\boldsymbol{\theta}})$ ;

```

---

The csIMI method fits the imputation model using the CSL method.

$$\hat{\boldsymbol{\alpha}}_{\text{csi}} = \arg \min_{\boldsymbol{\alpha}} \tilde{\mathcal{L}}(\boldsymbol{\alpha}),$$

where  $\bar{\boldsymbol{\alpha}}$  is chosen as the local solution at the central site.

$$\bar{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha}} \mathcal{L}_1(\boldsymbol{\alpha}).$$

Following the asymptotic property of the CSL estimator<sup>23</sup>, the covariance matrix of  $\hat{\boldsymbol{\alpha}}_{\text{csi}}$  is consistently estimated by

$$\text{Cov}(\hat{\boldsymbol{\alpha}}_{\text{csi}}) = \frac{1}{N_c} \nabla^2 \mathcal{L}_1(\boldsymbol{\alpha})^{-1} \Big|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_{\text{csi}}}.$$

### Algorithm 4

CSL MI algorithm

---

```

1 Find the estimate  $\bar{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha}} \mathcal{L}_1(\boldsymbol{\alpha})$ , which is the optimal estimate at site 1;
2 Send  $\bar{\boldsymbol{\alpha}}$  to all sites and receive  $\nabla \mathcal{L}_k(\boldsymbol{\alpha})|_{\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}}}$  back;
3 Find  $\hat{\boldsymbol{\alpha}}_{\text{csi}} = \arg \min_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha})$  and  $\text{Cov}(\hat{\boldsymbol{\alpha}}_{\text{csi}}) = \frac{1}{N_c} \nabla^2 \mathcal{L}_1(\boldsymbol{\alpha})^{-1} \Big|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_{\text{csi}}}$ ;
4 Sample  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_M$  independently from  $\mathcal{N}(\hat{\boldsymbol{\alpha}}_{\text{csi}}, \text{Cov}(\hat{\boldsymbol{\alpha}}_{\text{csi}}))$ ;
5 Send  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_M$  to all sites;
6 for  $m \leftarrow 1$  to  $M$  do
7   for  $k \leftarrow 1$  to  $K$  do Impute the missing data at site  $k$  based on  $\boldsymbol{\alpha}_m$ ;
8   Fit the analysis model and obtain  $\hat{\boldsymbol{\theta}}_m$  and  $\text{Cov}(\hat{\boldsymbol{\theta}}_m)$ ;
9 end
10 Combine the results by Rubin’s rule to obtain  $\hat{\boldsymbol{\theta}}$  and  $\text{Cov}(\hat{\boldsymbol{\theta}})$ ;

```

---

The aforementioned algorithms can be used for a wide range of imputation models including, but not limited to, GLMs. But, when applied to a linear imputation model, the algorithms need to be designed to draw the error variance parameter  $\tau^2$  in different ways. The iMI and siMI methods can follow the procedure in (3) without modification. However, a direct application of the AVGM and CSL approaches would sample the error variance  $\tau^2$  from Gaussian, which does not ensure its positiveness. To address this issue, we propose the following alternative procedures for sampling  $\tau^2$ .

For avgmMI, we use  $SSE = \sum_k SSE^{(k)}$  where  $SSE^{(k)}$  is the sum of squared errors from site  $k$ .

$$SSE^{(k)} = \| \mathbf{x}_{1,c}^{(k)} - \mathbf{Z}_c^{(k)} \hat{\boldsymbol{\alpha}}^{(k)} \|_2^2.$$

The error variance is sampled from  $\tau^2 \sim \mathcal{IG}(N_c/2, SSE/2)$  and  $\boldsymbol{\alpha}$  is sampled from the multivariate Gaussian distribution with mean  $\hat{\boldsymbol{\alpha}}_{avgm}$  and the variance

$$\text{Cov}(\hat{\boldsymbol{\alpha}}_{avgm} | \tau^2) = \frac{\tau^2}{N_c^2} \sum_k n_c^{(k)2} (\mathbf{Z}_c^{(k)T} \mathbf{Z}_c^{(k)} + \lambda \mathbf{I})^{-1}.$$

Each site needs transfer  $\hat{\boldsymbol{\alpha}}^{(k)}$ ,  $SSE^{(k)}$ , and  $(\mathbf{Z}_c^{(k)T} \mathbf{Z}_c^{(k)} + \lambda \mathbf{I})^{-1}$  to the central site after fitting the local imputation model.

The cslMI method also follows the procedure in (3). It entails sampling  $\tau^2$  with  $SSE = N_c \| \mathbf{x}_{1,c}^{(1)} - \mathbf{Z}_c^{(1)} \hat{\boldsymbol{\alpha}}_{csl} \|_2^2 / n_c^{(1)}$ , which is based on the asymptotic variance of  $\hat{\boldsymbol{\alpha}}$  defined in Eq. (13) in ref. 23 and does not require additional communication between sites to compute, and then sampling  $\boldsymbol{\alpha}$  from the multivariate Gaussian distribution with mean  $\hat{\boldsymbol{\alpha}}_{csl}$  and the variance

$$\text{Cov}(\hat{\boldsymbol{\alpha}}_{csl} | \tau^2) = \frac{\tau^2 n_c^{(1)}}{N_c} (\mathbf{Z}_c^{(1)T} \mathbf{Z}_c^{(1)} + \lambda \mathbf{I})^{-1}.$$

In finite samples, the above-mentioned asymptotic approximation of SSE may contribute to deteriorating performance of the cslMI when sample size decreases, as evidenced in our numerical experiments.

**Distributed MI for general missing data patterns.** Since it is commonly encountered in practice to have multiple variables with missing values, it is of particular interest to develop privacy-preserving distributed MI methods for general missing data patterns.

A very popular method for handling general missing data patterns is the MI by chained equation, known as MICE in short<sup>20</sup>. Without loss of generality, we assume that the first  $q$  ( $q < p$ ) covariates, i.e.,  $(X_1, \dots, X_q)$ , have missing values. The MICE algorithm starts with an initial imputation. For example, the missing values of  $X_j$  can be imputed by random samples from the observed data. Then, for each  $j = 1, \dots, q$ , the missing values of  $X_j$  are imputed by a MI method, assuming that the imputed values of all the other variables were actually observed. A sweep of imputations for all  $q$  missing variables form an iteration, and multiple iterations are fulfilled until the distribution of imputed values becomes stationary before the first imputed dataset is sampled. Multiple iterations are carried out between each imputed dataset to alleviate autocorrelations. A total of  $M$  imputed datasets are collected and used to fit the analysis model separately, and the results are combined by the Rubin's rule. Readers are referred to ref. 20 for more details about MICE.

### Algorithm 5

Privacy-preserving MI by chained equation algorithm

---

```

1  for  $j \leftarrow 1$  to  $q$  do
2    Impute the missing values of  $X_j$ ,
3  end
4  for  $m \leftarrow 1$  to  $M$  do
5    repeat
6      for  $j \leftarrow 1$  to  $q$  do
7        Fit the imputation model for  $X_j$  using the samples for which  $X_j$  are
        observed and impute the missing values of  $X_j$  by iMI, avgmMI, cslMI,
        or siMI;
8      end
9    until multiple times;
10   Fit the analysis model and obtain  $\hat{\boldsymbol{\theta}}_m$  and  $\text{Cov}(\hat{\boldsymbol{\theta}}_m)$ ;
11 end
12 Combine the results by Rubin's rule to obtain  $\hat{\boldsymbol{\theta}}$  and  $\text{Cov}(\hat{\boldsymbol{\theta}})$ ;

```

---

In parallel with the distributed MI methods, we consider four privacy-preserving distributed MICE approaches, namely, iMICE, avgmMICE, cslMICE, and siMICE. The generic privacy-preserving MICE algorithm is summarized in Box 5. Unlike the distributed MI methods for the univariate missing pattern, each imputation model is fitted multiple times in MICE. Therefore, these distributed

MICE methods may not be communication-efficient except iMICE, which requires no communication between each site and the central site.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

The real data analyzed in this article were provided by the Georgia Coverdell Acute Stroke Registry and restrictions apply to the availability of these data. Request for access to the data should be submitted to and approved by the Georgia Coverdell Acute Stroke Registry.

### Code availability

We used R software (Version 3.6.3) including custom code and existing R package (mice) to conduct the simulations and real data analysis. All relevant R codes are publicly available from <https://github.com/changee/MIDist>.

Received: 15 April 2020; Accepted: 2 October 2020;

Published online: 29 October 2020

### References

- Naveed, M. et al. Privacy in the genomic era. *ACM Comput. Surv.* **48**, 6:1–6:44 (2015).
- Jiang, X., Sarwate, A. D. & Ohno-Machado, L. Privacy technology to support data sharing for comparative effectiveness research: a systematic review. *Med. Care* **51**, S58 (2013).
- Homer, N. et al. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
- Brakerski, Z. Fully homomorphic encryption without modulus switching from classical gapsvp. in *Advances in Cryptology-CRYPTO 2012*, (Safavi-Naini, R. and Canetti, R. (eds)) 868–886 (Springer, 2012).
- Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
- Wang, R., Li, Y. F., Wang, X., Tang, H. & Zhou, X. Learning your identity and disease from research papers: information leaks in genome wide association study. in *Proceedings of the 16th ACM conference on Computer and communications security*, 534–544 (ACM, 2009).
- Brown, J. S. et al. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Medical Care* **48**, S45–S51 (2010).
- Kahn, M. G. et al. Transparent reporting of data quality in distributed data networks. *eGems* **3**, 7 (2015).
- Weeks, J. & Pardee, R. Learning to share health care data: A brief timeline of influential common data models and distributed health data networks in u.s. health care research. *eGEMS* **7**, 4 (2019).
- Ohno-Machado, L. et al. pscanner: Patient-centered scalable national network for effectiveness research. *J. Am. Med. Inform. Assoc.* **21**, 621–626 (2014).
- Toh, S., Platt, R., Steiner, J. F. & Brown, J. S. Comparative-effectiveness research in distributed health data networks. *Clin. Pharmacol. Ther.* **90**, 883–887 (2011).
- Davies, M., Erickson, K., Wyner, Z. & Malenfant, J. M. Software-enabled distributed network governance: The popmednet experience. *eGEMS* **4**, 5 (2016).
- Kantarcioglu, M. A survey of privacy-preserving methods across horizontally partitioned data. in *Privacy-Preserving Data Mining*, (Aggarwal, Charu C. and Yu, Philip S (eds)), 313–335 (Springer, 2008).
- Shortreed, S. M., Cook, A. J., Coley, R. Y., Bobb, J. F. & Nelson, J. C. Challenges and opportunities for using big health care data to advance medical science and public health. *Am. J. Epidemiol.* **188**, 851–861 (2019).
- Wells, B. J., Chagin, K. M., Nowacki, A. S. & Kattan, M. W. Strategies for handling missing data in electronic health record derived data. *eGEMS* **1**, 1035 (2013).
- Penny, K. I. & Atkinson, I. Approaches for dealing with missing data in health care studies. *J. Clin. Nurs.* **21**, 2722–2729 (2012).
- Little, R. J. & Rubin, D. B. *Statistical Analysis With Missing Data* (John Wiley & Sons, 2014).
- Rubin, D. *Multiple Imputation for Nonresponse in Surveys*. (Wiley, New York, 1987).

19. Raghunathan, T. E. & Siscovick, D. S. A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Appl. Stat.* 335–352 (1996).
20. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67 (2011).
21. Jagannathan, G. & Wright, R. N. Privacy-preserving imputation of missing data. *Data Knowl. Eng.* 65, 40–56 (2008).
22. Zhang, Y., Duchi, J. C. & Wainwright, M. J. Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.* 14, 3321–3363 (2013).
23. Jordan, M. I., Lee, J. D. & Yang, Y. Communication-efficient distributed statistical inference. *J. Am. Stat. Assoc.* 114, 668–681 (2019).
24. Wood, A. et al. Differential Privacy: A Primer for a Non-Technical Audience. *Vanderbilt Journal of Entertainment & Technology Law* 21, 209 (2018).
25. Xiao, Y., Xiong, L., Fan, L., Goryczka, S. & Li, H. Dpcube: differentially private histogram release through multidimensional partitioning. *Trans. Data Priv.* 7, 195–222 (2014).

### Acknowledgements

This work is partly supported by NIH grants R01GM124111. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

### Author contributions

All authors participated in the design of the methods. C.C. implemented the methods in R and conducted the experiments. C.C., Y.D., and Q.L. wrote the paper. Q.L. supervised the research. All authors reviewed and revised the paper.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-19270-2>.

**Correspondence** and requests for materials should be addressed to Q.L.

**Peer review information** *Nature Communications* thanks Samprit Banerjee, Jeffrey Brown and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020