SynTL: A synthetic-data-based transfer learning approach for multi-center risk prediction

Tian Gu, PhD[1] and Rui Duan, PhD[1*]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
*corresponding author: rduan@hsph.harvard.edu

Abstract

Objectives: We propose a synthetic-data-based transfer learning approach (SynTL) to incorporate multi-center healthcare data for improving the risk prediction of a target population, accounting for challenges including heterogeneity, data sharing, and privacy constraints.

Methods: SynTL combines locally trained risk prediction models from each source population with the target population to adjust for the data heterogeneity through a flexible distance-based transfer learning approach. Heterogeneity-adjusted synthetic data are then generated for source populations where individual-level data are not shareable. The synthetic data are then combined with the target and source data for joint training of the target model. We evaluate SynTL via extensive simulation studies and an application to multi-center data from the electronic Medical Records and Genomics (eMERGE) Network to predict extreme obesity.

Results: Simulation studies show that SynTL outperforms methods without adjusting for the data heterogeneity and methods that are trained in a single population over a wide spectrum of settings. SynTL has low communication costs where each participating site only needs to share parameter estimates to the target, requiring only one round of communication. SynTL protects against negative transfer when some source populations are highly different from the target. Using eMERGE data, SynTL achieves an area under the receiver operating characteristic curve (AUC) around 0.79, which outperforms other benchmark methods (0.50 - 0.67).

Conclusion: SynTL improves the risk prediction performance of the target population, and is robust to the level of heterogeneity between the target and source populations. It protects patient-level information and is highly communication efficient.

Keywords: data privacy, multicenter study, risk prediction, synthetic data, transfer learning

Introduction

Risk prediction models play an important role in precision health research and practice, and have shown great potential to assist clinical decision-making and enhance the quality of care delivered to patients[19]. In recent years, growing efforts have been devoted to the construction of institutional or national biobanks, where individuals' electronic health records (EHR) are linked with genomics, imaging, and behavioral observations. For example, the UK Biobank is a prospective cohort of over 500,000 samples with both genetic data and health information[45]. In the United States, institutional biobanks such as Bio*Me*[1,5] at the Icahn School of Medicine at Mount Sinai, BioVU at Vanderbilt University[38,43], and Mass General Brigham (MGB) Biobank[26] enroll increasing number of participants every year to measure their genetic variants, biomarkers, metabolic data and a large number of disease phenotypes. These datasets are valuable resources to assess individuals' risks of many complex diseases.

Despite the availability of multiple biobanks, one common phenomenon is the diminished performance when applying models trained in one dataset to a different dataset, which is known as the lack of model transferability and portability[20]. The low portability may be explained by the heterogeneity in the underlying distribution of data across different populations. For example, compared to many healthcare center-based biobanks, the UK Biobank has older participants with lower disease prevalence across many disease categories[4]. Another example is the lack of transferability of polygenic risk scores across ethnic groups[16,33], resulting from the substantial heterogeneity in the genetic architectures, linkage disequilibrium, and allele frequencies[16,33].

Practically, it is often of interest to construct a risk prediction model in a particular population (referred to as the target population hereafter), for example, a demographic sub-population or one of the healthcare centers in a research network. If substantial data are available from the target population, risk prediction models trained using only the target data could reach desired performance. However, in many cases, only limited amount of data are observed from the target population. For example, the target population of interest may be an underrepresented demographic

sub-population or a healthcare center with very few labeled data. In such cases, training a risk prediction model using only the target data may lead to unsatisfied performance due to the limited sample size, and therefore borrowing information from other existing data sources becomes much more necessary. Methodologically, it is still a challenge to effectively and efficiently compile available data sources towards risk prediction in a target population, while addressing practical issues such as data sharing constraints and heterogeneity across populations.

Protected by privacy rules and regulations, individual-level data are less feasible to be shared across institutions[6,10,12]. On the contrary, summary statistics, such as parameter estimates, sample mean, or quantile statistics are considered less sensitive to be shared across sites[53]. Consequently, distributed or federated methods based on summary statistics are developed for jointly training models across multiple datasets without sharing individual-level data. For example, a series of algorithms were developed for regression models by sharing gradients of objective functions[13,14,32,39,54]. Other approaches involve sharing parameter estimates that are fitted locally at each site[9,24,27,58]. However, these work requires data from each site to be homogeneously distributed, which fails to account for potential heterogeneity. In computer science literature, federated learning approaches are mainly developed for collecting and updating models from users' devices, where user-specific model parameters are allowed to account for differences of each user Prayitno et al.[37]. Nevertheless, these approaches often require iterative model updating across devices and the central server[31,32,44], whereas in research networks, it is important to minimize the communication costs (especially the communication rounds) since obtaining and sharing data both involve extensive efforts and resources[3,21].

Transfer learning methods use different yet related models to improve the prediction of a target model[8,30,36,52], and is especially suitable for handling situations where the target population has a relatively limited sample size. Methods that combine transfer learning and federated learning can be highly relevant in the application of multi-biobank risk prediction. For example, Li et al.[29] proposed a federated transfer learning approach to leverage source data to the target population. Cai et al.[7] proposed an integrative estimation procedure for multi-task high-dimensional regression.

The most significant drawback of these methods is that they all require each site to share both the first and second-order derivatives of the loss function with respect to the model parameters. In a moderate or high dimensional setting, when the number of parameters (p) can be comparable or even greater than the sample size of a population (n), sharing second-order information ($O(p^2)$ numbers) can be challenging regarding privacy, data storage and communication cost[29].

In this work, we propose a transfer learning framework for high-dimensional risk prediction in a target population while incorporating heterogeneous source datasets. Our work contributes to the current literature from four perspectives. First, our method covers a general class of prediction models and a broad spectrum of situations where the source data may or may not be directly shareable with the target population. Second, we develop a synthetic data approach, such that participating sites (biobanks) only need to transfer fitted model parameters to the target population if individual-level data are not allowed to be transferred, which protects patient privacy. Compared to current federated transfer learning methods discussed in Cai et al.[7], Li et al.[29], our method has substantially low communication costs, especially when the dimension of predictors is high. Third, the proposed method adjusts for the heterogeneity across populations by allowing population-specific model parameters. Our method avoids negative transfer from sites that are potentially highly different from the target population using a data-driven way, and therefore it is robust to the level of heterogeneity across populations[49,51]. Lastly, the proposed method has improved risk prediction accuracy compared to many benchmark methods based on summary-level statistics, and can reach comparable performance to the transfer learning methods that require direct pooling of individual-level data.

Methods

Notations and problem setup

We consider the setting where we have $K + 1$ datasets collected from $K + 1$ populations, respectively. We denote the target population by $k = 0$, and the K source populations by $k = 1, \ldots K$. We consider a general situation, where among the K source datasets, M can be directly combined

with the target dataset, and the remaining $K - M$ datasets have data-sharing constraints that the individual-level data are not allowed to be shared. Note that $M = 0$ indicates a completely federated setting, and $M = K$ indicates a centralized setting where all data are stored together. The k-th site has sample size n_k, and the total sample size is $N = \sum_{k=0}^{K} n_k$.

For the k-th population, we observe the outcome variable $Y_k \in \mathbb{R}^{n_k}$ and a set of p predictors $X_k \in \mathbb{R}^{n_k \times p}$. Consider a generic class of risk prediction models where fitting the model involves a loss function indexed by a set of parameters, denoted by $L(\boldsymbol{\beta}; Y, X)$. With a parameter estimate $\hat{\boldsymbol{\beta}}$, the predicted outcome is given by $\hat{y} = f(x, \hat{\boldsymbol{\beta}})$. For example, in a generalized linear models (GLM), $L(\cdot)$ is the negative log-likelihood function, and $\boldsymbol{\beta}$ corresponds to the regression coefficients[34]; when using support vector machine (SVM), $L(\cdot)$ is the Hinge loss function[11]. To account for the heterogeneity, we assume that the underlying population-level parameters are different across populations, i.e.,

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{\beta}} E\{L(\boldsymbol{\beta}; Y_k, X_k)\},$$

where the expectation is taken with respect to the distribution of data at the k-th population. The target parameter $\boldsymbol{\beta}_0$ can be different from $\boldsymbol{\beta}_k$ in the k-th source population, $k \in \{1, \ldots, K\}$. Our goal is to estimate $\boldsymbol{\beta}_0$ of the target population, using data from both the target population and K source populations, given the challenge that the distribution of data and the underlying parameters are different across populations.

The key of transfer learning is to characterize and leverage the similarities between the source and the target populations. Previous work such as Li et al.[29],[30], Tian and Feng[47], Xu and Bastani[55] characterize the similarity between the target model and the source models by a L_0 distance measure $d(\boldsymbol{\beta}_0, \boldsymbol{\beta}_k) = \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k\|_0$. A smaller distance implies fewer parameters have different values between the k-th source and the target populations, allowing more information to be borrowed from the k-th source population. Although such characterization is more flexible than some related work that requires pre-specifying which parameters are shared across populations and which are population-specific[15], it is still stringent as it assumes most of the parameters values to be identical

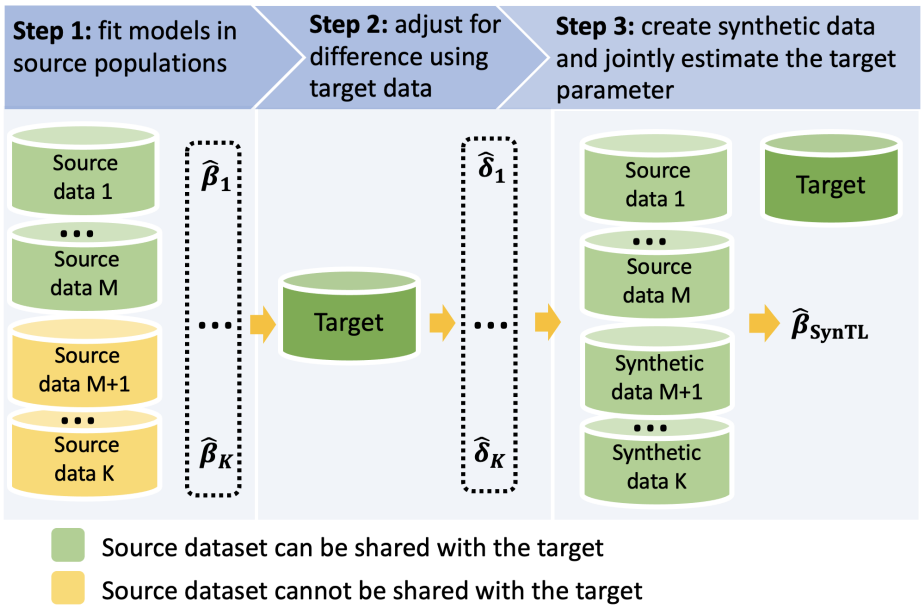

Figure 1: Workflow of SynTL.

across populations. To allow more flexibility, we propose a general distance measure defined as $d(\boldsymbol{\beta}_0, \boldsymbol{\beta}_k) = \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_0\|_q$, where q can take any positive numbers. Choosing $q = 0$ leads to the similarity measure considered in Li et al.[29,30], Tian and Feng[47], which quantifies how many parameters have identical values across two populations. When $q = 2$, it represents the Euclidean distance characterizing the overall differences over all parameters. In practice, one can flexibly choose q to better capture the differences in the model parameters across populations, which can be pre-specified or determined in a data-driven way through methods such as cross validation. In addition, weighted distance measures can be designed if prior knowledge is available regarding which parameters are more likely to be similar across two populations.

Proposed synthetic-data-based transfer learning (SynTL)

We propose a synthetic-data-based transfer learning (SynTL) approach, which involves three main steps (see Figure 1). To describe the method, we keep a generic form of the risk prediction model, and use a penalized logistic regression model as an example.

Step 1: In each source dataset, fit a risk prediction model and obtain the source-specific model parameters. The source-specific parameter estimate from the k-th site, denoted by $\hat{\boldsymbol{\beta}}^{(k)}$, can be

obtained through minimizing the loss function $L(\boldsymbol{\beta}_k; X_k, Y_k)$ determined by the risk prediction method. When additional assumptions are imposed on the model parameters, for example, sparsity or other structured patterns, additional penalty terms or constraints can be added to the optimization. A generic form is summarized as

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\beta_k} L(\boldsymbol{\beta}_k; X_k, Y_k) + \lambda_k \mathcal{P}(\boldsymbol{\beta}_k), \tag{1}$$

where $\mathcal{P}(\cdot)$ is some penalty term (added when necessary) and λ_k is a tuning parameter that can be selected through cross validation. Using a L_1 penalized logistic regression as an example, one can take $L(\boldsymbol{\beta}_k; X_k, Y_k) = \sum_{i=1}^{n_k} \log\{1 + \exp(x_{ik}^T \boldsymbol{\beta}_k)\} - y_{ik} x_{ik}^T \boldsymbol{\beta}_k$, where x_{ik} are the predictors including the intercept term for the i-th individual in the k-th site, y_{ik} is the outcome, and $\mathcal{P}(\boldsymbol{\beta}_k) = \|\boldsymbol{\beta}_k\|_1$. The K source estimators $\{\hat{\boldsymbol{\beta}}_k\}_{k=1}^K$ are then shared with the target.

Step 2: In the target site, adjust for the difference between $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_0$. With the target data, we construct the target loss function $L(\boldsymbol{\beta}_0; X_0, Y_0)$. If we do not want to borrow information from any source populations, we can get the *target-only* estimator by optimizing the target loss function, i.e.,

$$\hat{\boldsymbol{\beta}}_0 = \arg\min_{\beta_0} L(\boldsymbol{\beta}_0; X_0, Y_0) + \lambda_0 \mathcal{P}(\boldsymbol{\beta}_k). \tag{2}$$

When $d(\boldsymbol{\beta}_0, \boldsymbol{\beta}_k)$ is small, the source estimator $\hat{\boldsymbol{\beta}}_k$ can help reduce the search space of the target parameters. This is achieved through penalization on the difference between $\hat{\boldsymbol{\beta}}_k$ and $\boldsymbol{\beta}_0$. Specifically, denote $\boldsymbol{\delta}_k = \boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_k$, we can learn $\boldsymbol{\delta}_k$ through

$$\hat{\boldsymbol{\delta}}_k = \arg\min_{\delta_k} L(\hat{\boldsymbol{\beta}}_k + \boldsymbol{\delta}_k; X_0, Y_0) + \lambda_{0,k} Q(\boldsymbol{\delta}_k) \tag{3}$$

where $Q(\boldsymbol{\delta}_k)$ is a penalty term depending on the similarity measure $d(\boldsymbol{\beta}_0, \boldsymbol{\beta}_k)$ and $\lambda_{0,k}$ is a tuning parameter. For example, if it is believed that $d(\boldsymbol{\beta}_0, \boldsymbol{\beta}_k) = \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_0\|_0$ is small, we may choose $Q(\boldsymbol{\delta}_k) = \|\boldsymbol{\delta}_k\|_0$, which leads to a non-convex optimization[48]; or we may choose $Q(\boldsymbol{\delta}_k) = \|\boldsymbol{\delta}_k\|_1$, which is a convex relaxation of the L_0 penalty and also leads to sparse solutions[25]. With such penalization, the source estimator $\hat{\boldsymbol{\beta}}_k$ is leveraged to guide the learning of the target parameter $\boldsymbol{\beta}_0$.

Step 3: Jointly estimate $\boldsymbol{\beta}_0$ through a synthetic data approach. To jointly estimate $\boldsymbol{\beta}_0$, if all the individual-level data can be pooled together (i.e., $M = K$), we propose a pooled transfer learning

estimator, defined as

$$\hat{\boldsymbol{\beta}}_{\text{PooledTL}} = \arg\min_{\beta} \Big\{ \underbrace{L(\boldsymbol{\beta}; X_0, Y_0)}_{\text{target data}} + \sum_{k=1}^{K} \underbrace{L(\boldsymbol{\beta} - \hat{\boldsymbol{\delta}}_k; X_k, Y_k)}_{k\text{-th source data}} + \lambda_\beta \mathcal{P}(\boldsymbol{\beta}) \Big\}, \tag{4}$$

which we refer to as *PooledTL* hereafter. In this joint estimation, $\hat{\boldsymbol{\delta}}_k$ from Step 2 is used to adjust for the difference between the k-th source population and the target population.

When $M < K$, however, it is infeasible to obtain the PooledTL estimator since $K-M$ populations cannot share the individual-level data. To tackle this challenge, we propose a synthetic data method, where the idea is to generate $K-M$ sets of heterogeneity-adjusted synthetic data, each corresponding to a source population. These synthetic datasets will then be combined with the target dataset and the M source datasets to jointly estimate $\boldsymbol{\beta}_0$. For risk prediction, since we are interested in the conditional distribution $Y|X$, the marginal distribution of the synthetic data does not need to match the source population. Therefore, each synthetic dataset is generated such that the predictors share the same distribution as the target predictors, and the synthetic outcome variable is generated from the heterogeneity-adjusted source estimator, defined as $\hat{\boldsymbol{\beta}}_0^{(k)} = \hat{\boldsymbol{\beta}}_k + \hat{\boldsymbol{\delta}}_k$.

Specifically, without loss of generality, we assume that the first M source datasets, denoted by $k = 1, \ldots, M$, can share individual-level data with the target. For each $k \in \{M+1, \ldots, K\}$, we first generate the synthetic predictors by a random sampling from the target X_0, with sample size $n_k^{\text{syn}} = n_k \times r_k$ times, denoted as X_k^{syn}, where r_k is some positive integer. We will discuss the choice of r_k in the following paragraph. We then generate the synthetic outcome $\hat{Y}_k^{\text{syn}}$ from the specified risk prediction model with X_k^{syn} using $\hat{\boldsymbol{\beta}}_0^{(k)}$. For example, given a logistic regression, we generate $\hat{Y}_k^{\text{syn}}$ through a Binomial distribution with estimated probability $\hat{Pr}(Y_k^{\text{syn}}|X_k^{\text{syn}}) = expit(X_k^{\text{syn}T}\hat{\boldsymbol{\beta}}_0^{(k)})$, where $expit(x) = (1+e^{-x})^{-1}$. Finally, the SynTL estimator can be obtained through jointly analyzing a combined data of the target, the source and the synthetic data:

$$\hat{\boldsymbol{\beta}}_{\text{SynTL}} = \arg\min_{\beta} \Big\{ \underbrace{L(\boldsymbol{\beta}; X_0, Y_0)}_{\text{target data}} + \sum_{k=1}^{M} \underbrace{L(\boldsymbol{\beta} - \hat{\boldsymbol{\delta}}_k; X_k, Y_k)}_{\text{source data that can be shared}} + \sum_{k=M+1}^{K} \underbrace{L(\boldsymbol{\beta}; X_k^{\text{syn}}, \hat{Y}_k^{\text{syn}})}_{\substack{\text{heterogeneity-adjusted}\\ \text{synthetic data}}} + \lambda_\beta \mathcal{P}(\boldsymbol{\beta}) \Big\}. \tag{5}$$

A detailed algorithm summarizing steps 1 to 3 are shown in Algorithm 1. Similar ideas of creating

synthetic datasets to improve model performance while protecting individuals' privacy have been considered in association studies[23], survey methodologies[40,41,42] and causal inference[46]. Compared to their methods, our approach considers a high-dimensional risk prediction model and adjusts cross-site heterogeneity before creating synthetic data. Compared to existing federated learning methods that require iterative communication and second-order information, SynTL is much more communication-efficient as it only requires the model estimates from each site. Compared to the PooledTL approach, SynTL does not require access to individual-level data from source populations.

Algorithm 1: the SynTL algorithm

Data: Target dataset $\{X_0, Y_0\}$, source dataset $\{X_k, Y_k\}_{k=1}^{K}$, and replication number of the synthetic data r_k.

Result: $\hat{\boldsymbol{\beta}}_{\text{SynTL}}$

for $k = 1, \ldots, K$ **do**

  Compute $\hat{\boldsymbol{\beta}}_k$ from equation (1)

  Transmit $\hat{\boldsymbol{\beta}}_k$ to the target.

end

for $k = 1, \ldots, K$ **do**

  compute $\hat{\boldsymbol{\delta}}_k$ from equation (3)

end

for $k = M + 1, \ldots, K$ **do**

  Obtain $\hat{\boldsymbol{\beta}}_0^{(k)} = \hat{\boldsymbol{\beta}}_k + \hat{\boldsymbol{\delta}}_k$;

  Randomly sample $r_k \times n_k$ rows with replacement from X_0, denoted as X_k^{syn}.

  Generate the synthetic outcome $\hat{Y}_k^{\text{syn}}$ from X_k^{syn} using the specified risk prediction model with $\hat{\boldsymbol{\beta}}_0^{(k)}$.

end

Combine the target and the first M datasets, together with synthetic data and obtain $\hat{\boldsymbol{\beta}}_{\text{SynTL}}$ using equation (5).

The optimal choice of r_k depends on the similarity between the source and the target. As our numerical investigation suggests (See Figure S1 in the Supplementary Material), when the distance between the source and target populations is low or moderate, increasing r_k will improve the performance of the SynTL. When the difference is large, increasing r_k will decrease the performance. In both cases, increasing r_k will reduce the gap of the performance between SynTL and PooledTL, and the performance of SynTL will become relatively stable when r_k is large.

In practice, if computationally allowed, we suggest choosing a large r_k for all $k \in \{1, \ldots, K\}$, assuming all the source populations are helpful. As this assumption might not be true and some source populations might be highly different from the target, we introduce an additional step to prevent negative transfer in the following subsection.

A safeguard to prevent negative transfer

When some of the source populations are highly different from the target population, including the source estimates might negatively impact the performance of the SynTL method. Intuitively, when the distance $d(\boldsymbol{\beta}_k, \boldsymbol{\beta}_0)$ is large, estimating the difference $\boldsymbol{\delta}_k = \hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_0$ might be more difficult compared to directly estimating $\boldsymbol{\beta}_0$, which can diminish the estimation accuracy when including the k-th source population in the SynTL algorithm. It is essential to avoid "bad" sites that can worsen the performance. To achieve this, we propose to first rank the K sites by the performance of source estimators $\hat{\boldsymbol{\beta}}_k$ on the target training data. One straightforward way to avoid negative transfer is to only include sites with performance higher than a certain threshold. The threshold can be either pre-specified or determined by cross-validation. Another way is to sequentially incorporate populations based on their rankings. For example, we set the first estimator as the target-only estimator $\hat{\boldsymbol{\beta}}_0$, which does not include any source population. We can then only include the site whose model has the best performance on the target training data, and apply Algorithm 1 to obtain $\hat{\boldsymbol{\beta}}^{(1)}_{\text{SynTL}}$. Sequentially including more source populations, we can further obtain $\hat{\boldsymbol{\beta}}^{(2)}_{\text{SynTL}}, \ldots, \hat{\boldsymbol{\beta}}^{(K)}_{\text{SynTL}}$. A small validation dataset can then be use to evaluate the model with the best performance among these candidate models[28,50].

Validation of SynTL

Simulation studies

We evaluate the performance of SynTL using simulated data generated under different simulation settings. We consider the case where we have a target site and three source sites. We generate predictors of dimension $p = 2000$ from multivariate Gaussian distributions with mean 0 and a population-specific covariance matrix Σ_k for the k-th population. We choose Σ_k to be a block-wise

matrix with 20 blocks each has dimension 100×100. We set all 20 blocks to be the same, denoted by B_k, with an auto-regressive structure such that $B_{k,ij} = {b_k}^{|i-j|}$. We randomly generate b_k from a Uniform distribution between $(0.2, 0.8)$.

With the simulated predictors, we then generate binary outcome through a logistic regression model with different regression coefficients, i.e., for the k-th population, the outcome is generated from $Pr(Y|X) = expit(X^T \boldsymbol{\beta}_k)$ for $k = 0, 1, 2, 3$. We assume the regression coefficient in the target population ($k = 0$) to be sparse with 100 non-zero effect sizes, which are generated from a Uniform distribution between $(-0.5, 0.5)$. To generate the regression coefficient $\boldsymbol{\beta}_k$ for the k-th source populations, we randomly select h_k entries of $\boldsymbol{\beta}_0$, in which we add a noise with magnitude 0.5 and a randomly assigned sign from $\{-1, 1\}$. By increasing h_k, we increase the number of parameters with different values from $\boldsymbol{\beta}_0$, representing the increasing heterogeneity between the k-th source and the target population.

We evaluate the performance of the methods by varying the level of heterogeneity (h_k) in each source populations, as well as the sample sizes n_k, with the target sample size $n_0 = 500$. For the the level of heterogeneity, we consider the following four settings:

(A) All three source sites have low-level heterogeneity, $h_1 = h_2 = h_3 = 20$;

(B) All three source sites have moderate level heterogeneity, $h_1 = h_2 = h_3 = 60$;

(C) All three source sites have high level heterogeneity, $h_1 = h_2 = h_3 = 120$;

(D) Two source source have low-level heterogeneity, $h_1 = h_2 = 20$, and one has high level heterogeneity, $h_3 = 120$.

We also consider two sample size settings:

1. $n_1 = 1000$, $n_2 = 1500$, $n_3 = 2000$;

2. $n_1 = 2000$, $n_2 = 3000$, $n_3 = 4000$.

We apply Algorithm 1 with the loss function of logistic model. Given the sparse setting, we choose $\mathcal{P}(\cdot) = \mathcal{Q}(\cdot) = \|\cdot\|_1$. When implementing SynTL, we use $r_k = 25$. We compare the performance of the proposed method with another four comparisons: (i) target-only: estimator obtained from equation (2); (ii) source-only: applying the model in target-only method using each source data; (iii) direct aggregation: weighted average of target-only and source-only estimates where weights are proportional to the inverse of prediction errors evaluated in a validation dataset; and (iv) SynTL: the proposed synthetic-data-based transfer learning method, where we consider four different cases: all source data cannot be shared with the target ($M = 0$); only the first source datasets can be shared with the target ($M = 1$); the first two source datasets can be shared with the target ($M = 2$); all data can be pooled together ($M = 3$, PooledTL). For each method, we calculate the average area under the operating characteristic curve (AUC) over 200 iterations.

Application to the prediction of extreme obesity using data from eMERGE

We evaluate the performance of SynTL using multi-center data from the eMERGE network obtained through the database of Genotypes and Phenotypes (dbGaP), where EHR-derived phenotypes for 55,029 subjects from ten participating sites were linked with DNA samples from biorepositories[22]. Among all subjects, 73% of the participants are White, 20% are African American (AA), 1% are Asian, and 6% are Unknown. We evaluate the performance of SynTL by constructing risk prediction models for extreme obesity using demographic features and genotypic data.

The extreme obesity phenotyping algorithm at eMERGE classifies each subject into either “Case", “Control", or “Neither Case nor Control". We remove subjects with a “Neither Case nor Control" status and only include subjects with verified case-control status. Excluding three children’s hospitals, we initially obtain data from seven adults sites: Vanderbilt University, Marshfield Clinic, Mayo Clinic, Northwestern University, Geisinger Health System, Icahn School of Medicine at Mount Sinai (Mount Sinai for short), and Group Health Cooperative. Predictors include age, gender, the top principal components (PCs) obtained from the genotypes, as well as single nucleotide polymorphisms (SNPs) that are strongly associated with extreme obesity selected based on p-values

from a genome-wide association study (GWAS) and linkage disequilibrium. The screening steps of SNPs are described in the Supplementary Material.

We treat AA at Vanderbilt University as the target population, where a target-only fitting strategy might have limited performance with a sample size of $n_0 = 646$. We select source populations with larger sample sizes than the target population. The sample sizes of the qualified populations are listed in Table 1, including six data centers and two racial groups. Although at eMERGE, all individual-level data can be combined, we use it to evaluate the performance of SynTL in a federated setting and consider that individual-level data cannot be shared across institutions. In this application, we have $K = 6$ and $M = 1$, as data from the Vanderbilt White population can be directly combined with the target population.

Table 1: Sample size across data centers.

	Target	Source					
		1	2	3	4	5	6
	Vanderbilt University	Vanderbilt University	Marshfield Clinic	Mayo Clinic	Northwestern University	Geisinger Health System	Mount Sinai
White	/	1523	1497	1321	1305	1023	/
African American	646	/	/	/	/	/	1142

The target data are randomly split into training, validation and testing data, with an approximate ratio of 8:2:3. In addition to the target testing dataset, we use all 167 AA samples from Northwestern University as an external testing dataset excluded from the training stage to evaluate the portability of the fitted models. We apply Algorithm 1 with the loss function $L(\boldsymbol{\beta}_k; X_k, Y_k) = \sum_{i=1}^{n_k} \log\{1 + \exp(x_{ik}^T \boldsymbol{\beta}_k)\} - y_{ik} x_{ik}^T \boldsymbol{\beta}_k$ and $\mathcal{P}(\cdot) = \mathcal{Q}(\cdot) = \|\cdot\|_1$. Over 10 iterations by randomly selecting the training, validation and testing data, we present the result of averaged AUC on the internal and the external testing data, respectively. We compare SynTL to target-only, six source-only, direct aggregation and PooledTL introduced in the previous simulation section.

Results

Results for simulation studies

Figures 2 and 3 present AUC over 200 simulation replications of four scenarios, under two different sample size configurations, respectively. In Panel A-C of the two figures, the three source populations have the same h_k (level of heterogeneity). As expected, when the level of heterogeneity increases between the source and the target populations, the performance of the source-only estimators on the target data decreases. With low heterogeneity (Panel A), the performance of the source estimators increases with the sample size. With moderate to high level of heterogeneity, the performance of source estimators may decrease with the same size (Panels B and C). With low to moderate heterogeneity (Panels A and B), the source estimators have higher AUC than the target-only estimator, while in the high heterogeneity setting (Panels C), the source estimators have lower AUC than the target-only estimator.

The proposed methods have higher performance than the target and source only estimators, as well as the direct aggregation approach. The variation of SynTL performance across different M is relatively small. For example, in Panel A of Figure 2, SynTL with $M = 3$ has the highest average AUC around 0.83, while SynTL with $M = 0$ has AUC around 0.81. When the heterogeneity is high (Panel C), we observing a slightly decreasing trend from $M = 0$ to $M = 3$, indicating that having individual-level data does not lead to improved performance than just sharing model estimates.

Panel D shows that SynTL is robust to the case where some source populations are highly different from the target, and can incorporate information from source populations that are truly helpful. Comparing Figure 2 to Figure 3, we see that the performance of the SynTL increases when we have larger sample sizes from source populations with low or moderate level of heterogeneity.

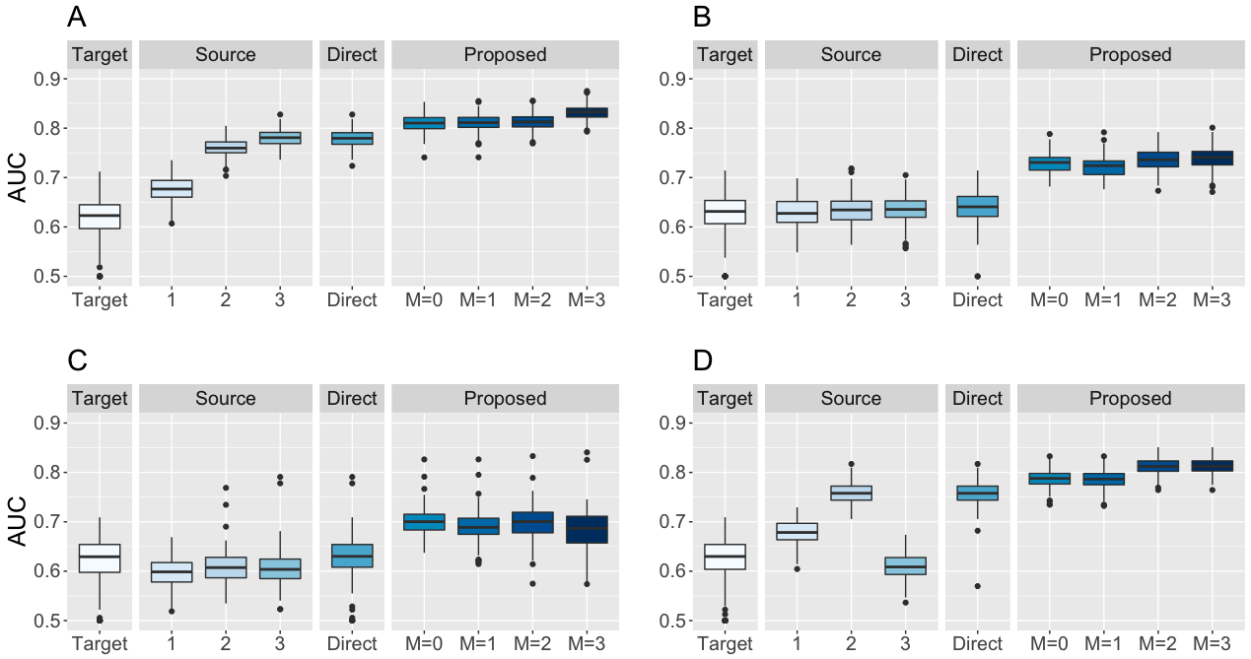

Figure 2: Area under the operating characteristic curve (AUC) over 200 simulation replicates with source sample sizes 1000, 1500, and 2000, respectively. Target: targe-only estimates; Source only: source-only estimates; Direct: direct weighted average of target-only and source-only estimates; Proposed: the SynTL methods with varying number of source datasets that can share individual-level data to the target (all source data cannot be shared with the target [$M = 0$]; only the first source datasets can be shared with the target [$M = 1$]; the first two source datasets can be shared with the target [$M = 2$]; all data can be pooled together [$M = 3$, PooledTL])

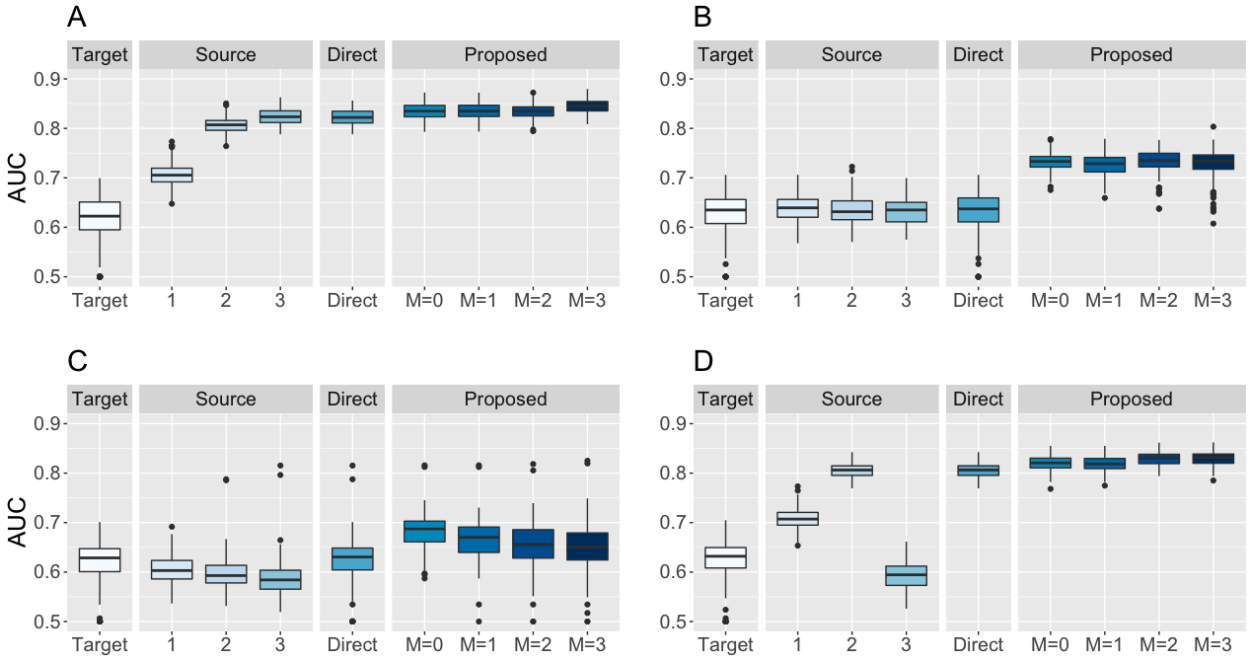

Figure 3: Area under the operating characteristic curve (AUC) over 200 simulation replicates with source sample sizes 2000, 3000, and 4000, respectively. Target: targe-only estimates; Source only: source-only estimates; Direct: direct weighted average of target-only and source-only estimates; Proposed: the SynTL methods with varying number of source datasets that can share individual-level data to the target (all source data cannot be shared with the target [$M = 0$]; only the first source datasets can be shared with the target [$M = 1$]; the first two source datasets can be shared with the target [$M = 2$]; all data can be pooled together [$M = 3$, PooledTL].

Results for predicting the risk of extreme obesity

After screening, 2047 SNPs are included in the model in addition to age, gender and top PCs. For the internal testing at Vanderbilt, as shown in Figure 4, the target-only estimator has an average AUC around 0.66. The six source estimators have average AUC ranging from 0.50 to 0.63, where the Geisinger model has the worst performance (with AUC of 0.50 indicating no prediction ability on the target data), and the Marshfield model has the best performance on the target population. All of the source estimators have lower performance compared to the target-only estimator. The direct aggregation by weighted averaging the target-only and source-only estimates show nearly identical performance compared to the target-only model. The proposed SynTL has an AUC around 0.78, which is comparable to PooledTL (AUC around 0.80), with an average 12% improvement compared to the target-only estimator. For external testing at Northwestern, the performance of

each method is relatively the same as being tested at Vanderbilt.

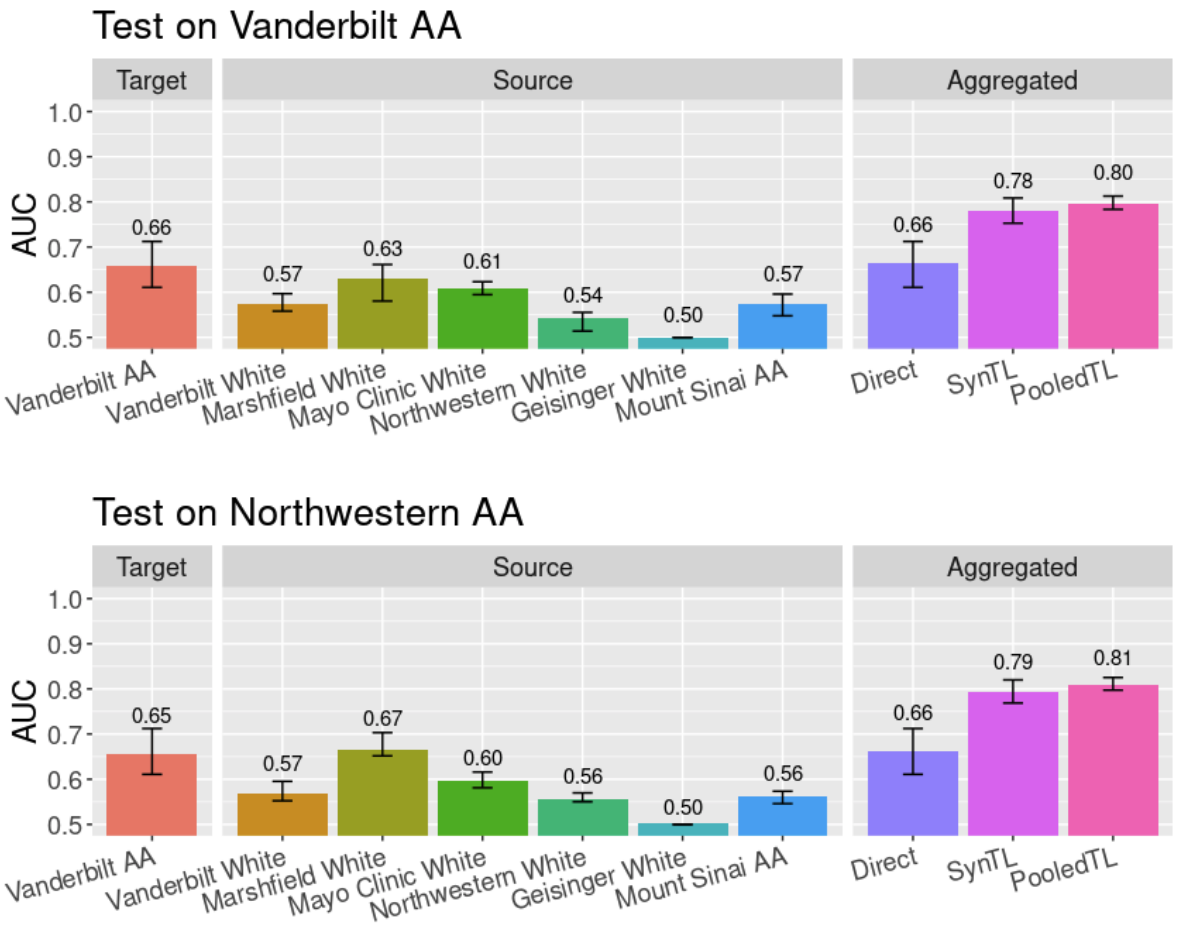

Figure 4: Area under the operating characteristic curve (AUC) of the compared methods for predicting extreme obesity using data from eMERGE. Target: model trained using data from target population only; Direct: weighted average of the target-only and the source-only models; SynTL: proposed method; PooledTL: transfer learning using all individual-level data. Each colored bar shows the average AUC over 10 random replications, the black error bars on top of each colored bar shows the 1st–3rd quantile of AUC.

Discussion

In this paper, we propose a transfer learning method that can be applied to both centralized and federated settings to integrate data from multiple populations for improving the risk prediction in a target population. Given the parameter estimates from the source populations, SynTL leverages the similarity across populations by a data-driven distance-based constraint, and adjust for the difference between the source and target parameters. Through a synthetic data approach, we then

generate heterogeneity-adjusted synthetic data corresponding to source data that can not be directly shared. Finally, we combine available source and target datasets with the synthetic datasets to jointly estimate the target parameters. Our SynTL algorithm can safeguard against source populations that are too different from the target, and it has low communication costs because only the fitted model parameters are required to be shared. Through numerical experiments and application to real-world multi-center data, we illustrate the validity and feasibility of SynTL for multi-center risk prediction.

One limitation of SynTL is that it requires all sites to observe the same set of predictors. In a multi-center study, each site likely observes different predictors, leading to additional heterogeneity in the data structure. SynTL is applicable in such cases if a joint feature space can be identified. Many methods existed for feature transformation in a non-federated setting[51]. However, federated feature transformation may still be challenging, especially when we want to avoid iterative communication across sites. Alternatively, missing features in a particular site could be first imputed from models trained at other sites[23,35,56]. However, data heterogeneity across sites and populations needs to be accounted for when applying imputation methods.

Our method protects individuals' privacy by only requiring parameter estimates across sites. This level of privacy protection is common in a collaborative research environment, where the risk for membership inference or tracing attack is low. In practice, when it is necessary, our method can be combined with other privacy protection techniques, for example, differential privacy[17,18], and data encryption[2,57] to further improve privacy protection. However, some of these approaches may compromise the prediction accuracy. Researchers need to balance the level of privacy protection with both the prediction performance and the communication cost in the design of the algorithm.

SynTL is an effective tool to improve prediction performance in a target population. The improvement is more significant when the target population has smaller sample size than the source populations. This indicates that SynTL can help reduce the gaps in the performance of prediction models across populations, which is caused by the limited representation of certain minority and disadvantaged groups. As the current biobank data heavily over-represents European ancestry

populations[16,33], it is a promising and meaningful future direction for us to study the utility of SynTL for improving the fairness of risk assessment tools in precision medicine.

Conclusion

SynTL can incorporate models trained from multiple healthcare institutions to improve the performance of risk prediction in a target population. SynTL is easy-to-implement and communication efficient, requiring each site to only share the fitted model parameters. It is robust to the level of heterogeneity across populations. SynTL is well-suited to be applied to multi-center studies to train robust risk prediction models and address the potential lack of representation of certain populations.

Conflict of Interest Statement

The authors (TG and RD) declare that there is no conflict of interest.

Data Availability Statement

The eMERGE Network data used in this work is publicly available through dbGaP phs000888.v1.p1 (https://ega-archive.org/studies/phs000888).

Supplementary Material

Investigation on the optimal choice of r_k

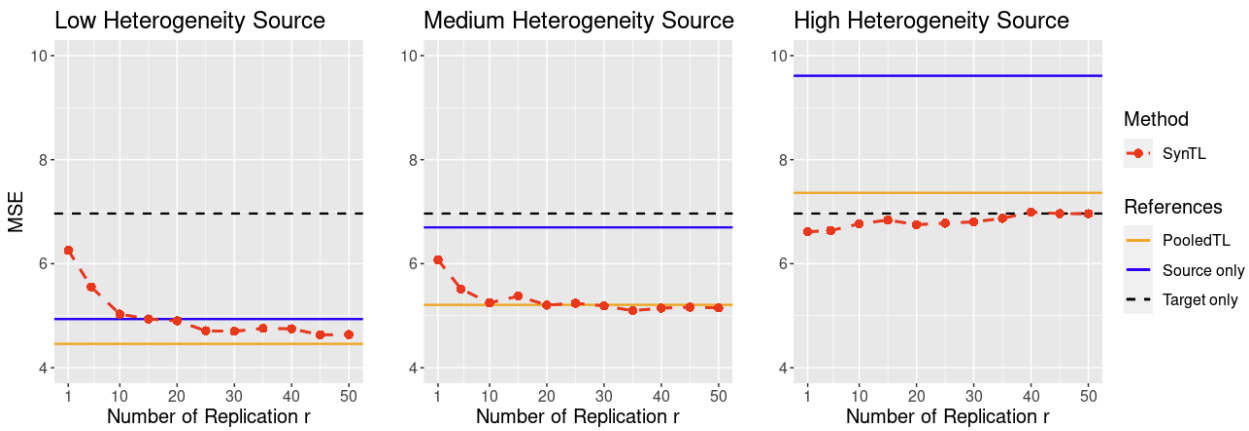

Figure S1: Mean squared error (MSE) of estimated coefficients over increasing synthetic data size from a single simulated dataset. One source population, i.e. K=1, is used as an example, varying heterogeneity level compared to the target population.

Screening of SNPs

To select the SNPs that are strongly associated with extreme obesity, we first perform genome-wide association study (GWAS) of extreme obesity using all the White data in eMERGE via PLINK software, treating the binary extreme obesity status as outcome and each of the SNPs as a covariate in a logistic regression adjusting for age, gender and the top principal components (PCs). Then we perform clumping and thresholding using eMERGE White data as linkage disequilibrium reference panel controlling for the between-SNP correlation to be less than 0.5 and p-value threshold at 0.005 (PLINK –clump-kb 500 –clump-p1 0.005 –clump-r2 0.5). The 2047 SNPs qualified above criteria are adjusted in the final model.

References

1 Abul-Husn N. S, Soper E. R, Odgis J. A et al. (2019). Exome sequencing reveals a high prevalence of brca1 and brca2 founder variants in a diverse population-based biobank. *Genome Medicine*, 12(1):2. PMCID: PMC6938627.

2 Acar A, Aksu H, Uluagac A. S et al. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4):1–35.

3 Agency E. C (2017). *Typical Cost Elements in Data-sharing*. Publications Office of the European Union.

4 Beesley L. J, Salvatore M, Fritsche L. G et al. (2020). The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Statistics in medicine*, 39(6):773–800.

5 Belbin G. M, Cullina S, Wenric S et al. (2021). Toward a fine-scale population health monitoring system. *Cell*, 184(8):2068–2083.e11. PMID: 33861964.

6 Bonomi L, Huang Y and Ohno-Machado L (2020). Privacy challenges and research opportunities for genomic data sharing. *Nature Genetics*, 52:646–654.

7 Cai T, Liu M and Xia Y (2021). Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, pages 1–15.

8 Cai T. T and Wei H (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128.

9 Chatterjee N, Chen Y.-H, P.Maas et al. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111:107–117.

10 Christen P, Ranbaduge T and Schnell R (2020). Linking sensitive data. *Methods and techniques for practical privacy-preserving information sharing. Cham: Springer*.

11 Cortes C and Vapnik V (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

12 Dove E. S and Phillips M (2015). *Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective*, pages 639–678. Springer International Publishing, Cham.

13 Duan R, Boland M. R, Moore J. H et al. (2018). Odal: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 30–41. World Scientific. PMCID: PMC6417819.

14 Duan R, Luo C, Schuemie M. J et al. (2020). Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association*, 27(7):1028–1036.

15 Duan R, Ning Y and Chen Y (2019). Heterogeneity-aware and communication-efficient distributed statistical inference. *arXiv preprint arXiv:1912.09623*.

16 Duncan L, Shen H, Gelaye B et al. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications*, 10(1):1–9.

17 Dwork C and Smith A (2010). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2).

18 Ficek J, Wang W, Chen H et al. (2021). Differential privacy in health research: A scoping review. *Journal of the American Medical Informatics Association*, 28(10):2269–2276.

19 Ginsburg G. S and Phillips K. A (2018). Precision medicine: from science to value. *Health Affairs*, 37(5):694–701.

20 Goldstein B. A, Navar A. M, Pencina M. J et al. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208.

21 Goodhill G. J (2014). Practical costs of data sharing. *Nature*, 509(33).

22 Gottesman O, Kuivaniemi H, Tromp G et al. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genetics in Medicine*, 15(10):761–771. PMCID: PMC3795928.

23 Gu T, Taylor J. M, Cheng W et al. (2019). Synthetic data method to incorporate external information into a current study. *Canadian Journal of Statistics*, 47(4):580–603.

24 Gu T, Taylor J. M and Mukherjee B (2021). A meta-inference framework to integrate multiple external models into a current study. *Biostatistics*, dio:10.1093/biostatistics/kxab017.

25 Hoerl E. A and Kennard W. R (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. DOI: 10.2307/1271436.

26 Karlson E. W, Boutin N. T, Hoffnagle A. G et al. (2016). Building the partners healthcare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations. *Journal of Personalized Medicine*, 6(1). PMCID: PMC4810381.

27 Kundu P, Tang R and Chatterjee N (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3):567–585.

28 Lecué G and Rigollet P (2014). Optimal learning with q-aggregation. *The Annals of Statistics*, 42(1):211–224.

29 Li S, Cai T and Duan R (2021). Targeting underrepresented populations in precision medicine: A federated transfer learning approach. *arXiv preprint arXiv:2108.12112*.

30 Li S, Cai T. T and Li H (2020). Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv preprint arXiv:2006.10593*.

31 Li W, Liu H, Yang P et al. (2016). Supporting regularized logistic regression privately and efficiently. *PloS one*, 11(6):e0156479.

32 Lu C.-L, Wang S, Ji Z et al. (2015). Webdisco: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6):1212–1219.

33 Martin A. R, Kanai M, Kamatani Y et al. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591. PMCID: PMC6563838.

34 Müller M (2000). *Generalized Linear Models*, pages 205–228. Springer Berlin Heidelberg, Berlin, Heidelberg.

35 Murray J. S (2018). Multiple imputation: a review of practical and theoretical findings. *Statistical Science*, 33(2):142–159.

36 Pan S. J and Yang Q (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

37 Prayitno C.-R. S, Karisma T. Putra H.-C. C, Tsai Y.-Y et al. (2021). A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications. *Applied Sciences*, 11(23):11191.

38 Pulley J, Clayton E, Bernard G. R et al. (2010). Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clinical and Translational Science*, 3(1):42–48. PMCID: PMC3075971.

39 R D, M R. B, Z L et al. (2019). Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385.

40 Raghunathan T. E, Reiter J. P and Rubin D. B (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1.

41 Reiter J. P (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531.

42 Reiter J. P and Kinney S. K (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, 28(4):583.

43 Roden D, Pulley J, Basford M et al. (2008). Development of a large-scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369. PMCID: PMC3763939.

44 Sarma K. V, Harmon S, Sanford T et al. (2021). Federated learning improves site performance in multicenter deep learning without data sharing. *Journal of the American Medical Informatics Association*, 28(6):1259–1264.

45 Sudlow C, Gallacher J, Allen N et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779–e1001779.

46 Tan X, Chang C.-C. H and Tang L (2021). A tree-based federated learning approach for personalized treatment effect estimation from heterogeneous data sources. *arXiv preprint arXiv:2103.06261*.

47 Tian Y and Feng Y (2021). Transfer learning under high-dimensional generalized linear models. *arXiv preprint arXiv:2105.14328*.

48 Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

49 Torrey L and Shavlik J (2010). *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global.

50 Tsybakov A. B (2014). Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 225–246.

51 Weiss K, Khoshgoftaar T. M and Wang D (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

52 Wiens J, Guttag J and Horvitz E (2014). A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706.

53 Wolfson M, Wallace S. E, Masca N et al. (2010). Datashield: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–1382.

54 Wu Y, Jiang X, Kim J et al. (2012). G rid binary lo gistic re gression (glore): building shared models without sharing data. *Journal of the American Medical Informatics Association*, 19(5):758–764.

55 Xu K and Bastani H (2022). Learning across bandits in high dimension via robust statistics. *arXiv:2112.14233v2*.

56 Xue F and Qu A (2020). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, pages 1–14.

57 Yi X, Paulet R and Bertino E (2014). Homomorphic encryption. In *Homomorphic encryption and applications*, pages 27–46. Springer.

58 Zhang H, Deng L, Schiffman M et al. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika*, 107(3):689–703.