Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

Dataset-to-Dataset Evaluation Before (and Without) Sharing Data

Keren Fuentes*[1], Mimee Xu*[2], Irene Y. Chen[3,4]

[1]Independent Researcher,
[2]New York University
[3]University of California, Berkeley
[4]University of California, San Francisco,
kerenfuentes313@gmail.com, mimee@nyu.edu, iychen@berkeley.edu,

Abstract

Privacy concerns and competitive interests impede data access for machine learning, due to the inability to privately assess external data's utility. This dynamic disadvantages smaller organizations that lack resources to aggressively pursue data-sharing agreements. In data-limited scenarios, not all external data is beneficial, and collaborations suffer especially in heavily regulated domains: metrics that aim to assess external data given a source e.g., approximating their KL-divergence, require accessing samples from both entities pre-collaboration, hence violating privacy. This conundrum disempowers legitimate data-sharing, leading to a false "privacy-utility trade-off". To resolve privacy and uncertainty tensions simultaneously, we introduce SecureKL, the first secure protocol for dataset-to-dataset evaluations with zero privacy leakage, designed to be applied preceding data sharing. SecureKL evaluates a source dataset against candidates, performing dataset divergence metrics internally with private computations, all without assuming downstream models. On real-world data, SecureKL achieves high consistency ($> 90\%$ correlation with non-private counterparts) and successfully identifies beneficial data collaborations in highly-heterogeneous domains (ICU mortality prediction across hospitals and income prediction across states). Our results highlight that secure computation maximizes data utilization, outperforming privacy-agnostic utility assessments that leak information.

Code — https://github.com/kere-nel/secure-data-eval
Extended version — https://arxiv.org/abs/2502.05765

1 Introduction

Since Hestness et al. (2017), empirical works have predetermined data as a key driver to performance gains, via the so-called "scaling laws" (Kaplan et al. 2020; Brown et al. 2020). Yet, accessing and combining datasets is persistently challenging. As datasets have evolved from small to larger (Sun et al. 2017), more diverse (Zhang et al. 2020), and more compute-optimal (Hoffmann et al. 2022), the field of machine learning continues to seek more data (Mahajan et al. 2018; Brown et al. 2020; Park et al. 2019; Reed et al. 2022; Sheller et al. 2020) and better ways to *combine* it

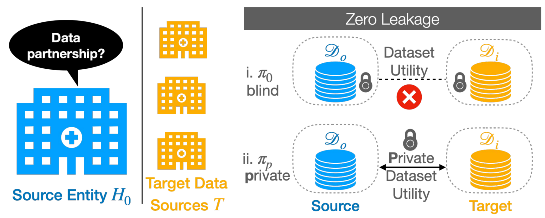

Figure 1: **Privacy can dis-incentivize data collaborations.** Without seeing external data, an organization has two strategies: **i**. **blind default** π_0: randomly selecting partnerships causes hesitation and hinders partnerships. **ii**. **private evaluation** π_p: securely assessing datasets *before* commitment.

(Shen et al. 2023; Nguyen et al. 2023; Petty, van Steenkiste, and Linzen 2024; Li et al. 2024; McKinney et al. 2020).

Strategically combining data from different sources promises enhanced models, but disempowers smaller organizations. While diverse, high quality data often improves performance, robustness, and fairness (Miller et al. 2021), access to such data significantly varies across entities and domains (Tenopir et al. 2011; L'Heureux et al. 2017). As domain-specific data becomes increasingly valuable (Alsentzer et al. 2019; Gururangan et al. 2020; Lee et al. 2020), data-owning entities are more reluctant to share it for free, opting instead to sell it in emerging data markets (Acemoglu et al. 2022; Huang et al. 2021; Liang et al. 2018). This dynamic disproportionately handicaps smaller organizations who lack both the resources to purchase data and the leverage to negotiate favorable sharing agreements.

Organizations may hesitate to commit to a potential partnership when unsure about the benefits. As Figure 1 illustrates, this "commitment issue" is not solely a privacy issue; it's the inability to privately assess an external dataset's utility *before* partnerships. This evaluation is, however, nontrivial. For example, adding in-domain data does not necessarily result in more performant models. Due to the precarious nature of domain shifts, machine learning models are sensitive to additional training data, often unpredictably when the source dataset is small (e.g., a single hospital's data) – a phenomenon known as "the dataset combination problem" (Koh et al. 2021; Wang et al. 2022; Taori et al. 2020; Bradley et al. 2021; Meng 2018; Shen, Raji, and Chen 2024).

*These authors contributed equally.

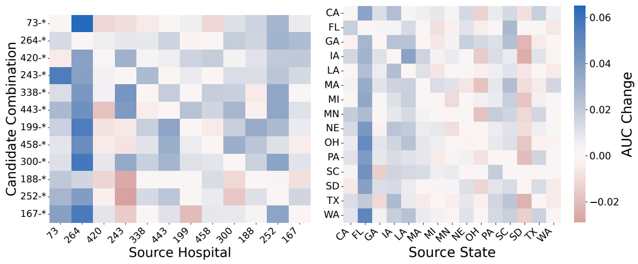

Figure 2: **The Dataset Combination Problem.** Real-world data collaborations are inherently uncertain. AUC change for a source entity, after incorporating external data across hospitals and states. **Left**: In eICU (Pollard et al. 2018), 10 out of 12 hospitals may see their mortality prediction model degrade for *some* potential hospital partners. **Right**: In Folktables (Ding et al. 2021), combining with random state leads to worse income prediction in 10 out of 15 states. (red is bad; exact values are reported in Appendix F)

Ideally, *all* target data should be considered to reduce uncertainty in costly data collaborations. Yet, **datasets owned by separate entities cannot be directly and fully accessed**, significantly limiting the practicality of non-private dataset measures (Shen, Raji, and Chen 2024; Ilyas et al. 2022).

Our work directly addresses this crux by recognizing both privacy and competitive incentives. First, *before* committing to acquiring unseen data, we enable organizations to privately gauge the relative utility of candidate datasets. Second, we provide strong privacy guarantees required of entities operating under stringent regulations e.g. healthcare providers to navigate data acquisition.

Developers are often uncertain about the most effective model before more data becomes available. This renders a secure data appraisal stage by Xu, Hannun, and Van Der Maaten (2022), which requires model parameters, not applicable. In this opportunistic setting, we ask:

Can we ascertain the differential utility of prospective datasets, without knowing the final model?

We introduce SecureKL, **the first divergence-based dataset measure with zero privacy leakage**. Our key insight is that private divergence computations (via secure multiparty computation (MPC)) are more data-efficient than sharing samples, thus achieving high performance. By preserving performance, SecureKL presents a compelling guarantee: **for both parties, privacy is fully protected while data utilization is maximized**.

Contributions A novel secure dataset-to-dataset evaluation protocol SecureKL (SKL) that reduces uncertainty in data utility under limited data and budget, producing privacy-preserving measures while using the maximum available samples. SKL achieves a $> 90\%$ correlation with privacy-violating counterparts across two real-world heterogeneous domains. Empirically, on ICU mortality prediction, SKL reliably chooses beneficial hospital(s) to partner with, outperforming data-leaking alternatives, including using demographic summaries or sharing data subsets.

Impact We provide a practical solution for organizations seeking to leverage collective data resources while maintaining privacy and competitive advantages. Our major advantage lies in reliability, especially when small organizations cannot afford to invest in detrimental partnerships. These results demonstrate the potential for wider data collaboration to advance machine learning applications in high-stakes domains while promoting more equitable access to data. Our code is available, and can be readily deployed to demonstrate potential data value preceding collaborations.

2 Background

2.1 Dataset Combination Problem

In high-stakes domains, incorporating additional datasets may *degrade* the model. In healthcare scenarios, both Compton et al. (2023) and Shen, Raji, and Chen (2024) showed that blindly acquiring new datasets can degrade model performance. We extend these findings by replicating the effect under our own experimental setup. As shown in Figure 2, combining data from one hospital or state with an external source often leads to inconsistent, and sometimes negative, changes in model performance. This non-monotonic behavior highlights the need to evaluate data partners *before* embarking on a full-fledged collaboration.

2.2 Trading Off Sample Utility For "Privacy"

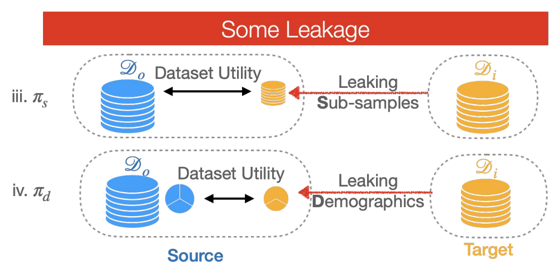

Figure 3: **Non-private evaluation strategies**. **iii**. **sub sampling** π_s: a subset of the target's data is shared. **iv**. **demographic summaries** π_d: the target entity discloses distributions by protected attributes, i.e. age, gender, or race.

Before an agreement, informed negotiations become impossible when entities do not expose data. In practice, data owners frequently resort to tiny samples (Centers for Medicare & Medicaid Services 2025) or summary statistics (e.g., race, gender, age) for making decisions (Figure 3). Yet, in data-limited settings, model performance is highly sensitive to new input. These heuristics are fickle, as sparse traits or limited samples often fail to capture the entire dataset's nuance and complexity, especially in heterogeneous domains, creating a perceived privacy-utility trade-off. We will show that secure computation, on the other hand, can avoid this tradeoff by measuring over *entire* datasets while maintaining privacy.

3 Problem Setup

Consider a binary prediction task for ICU patient mortality based on electronic medical records. A source hospital H_o has historical patient data $\mathcal{D}_o$ containing static past patient characteristics, prior medical records, and ICU outcomes. Other hospitals $\{H_i\}$ each has their patient data: $\{\mathcal{D}_i \mid i \in [1..N]\}$.

For this binary prediction task, hospitals typically optimize for performance metrics, for example the area under the receiver-operating characteristic curve (AUC). Using only their data, H_o can train a model $\mathcal{M}$ with parameters θ to achieve:

$$\mathrm{AUC}_o = \max_{f(\theta)} \mathrm{AUC}(\mathcal{M}, \mathcal{D}_o) \qquad \text{(Baseline AUC)}$$

where f is their chosen algorithm with parameter θ.

When H_o has exhausted their own internal data, they may benefit from incorporating additional target data sources $T \subset [1..N]$. By combining datasets, i.e., $\mathcal{D}_T = \{\mathcal{D}_i \mid i \in T\} \cup \mathcal{D}_o$, H_o can potentially achieve better results:

$$\mathrm{AUC}_T = \max_{f(\theta)} \mathrm{AUC}(\mathcal{M}, \mathcal{D}_T). \qquad \text{(Combined AUC)}$$

We define the potential improvement from data addition as $\delta_T = \delta_{(o,T)} = \mathrm{AUC}_T - \mathrm{AUC}_o$. To add a single additional data source by setting $T = \{i\}$, the improvement is $\delta_i = \delta_{(o,i)} = \mathrm{AUC}_i - \mathrm{AUC}_o$. This leads to our central question:

Without seeing target data, how does a hospital ascertain potential data sources to combine with?

Formally, given $n \leq N$, we seek a strategy π that selects n target datasets $T = \pi(\mathcal{D}_o, n)$ to maximize model utility:

$$\pi^*(\mathcal{D}_o, n) = \underset{T \subset \binom{[1...N]}{n}}{\arg\max}\ \mathrm{AUC}_T \qquad \text{(Ideal Combination)}$$

Practical Considerations. Computing every subset $T \subset \binom{[1...N]}{n}$'s associated δ_T is exponential in n. To make this problem tractable, we make two key assumptions. First, we apply strategies greedily, selecting top-ranked target datasets. With the ultimate objective of improving the source hospital's prediction task, we fix H_o; to compare the tradeoffs between strategies in Section 4, we apply each π greedily to select top-n institution(s) for H_o without replacement. Second, in data constrained settings, we aim to maximize the probability of positive improvement: $P_{H_o \sim \mathbf{H}}(\delta_T > 0)$.

Kullback–Leibler Divergence. Our approach uses Kullback-Leibler (KL)-divergence-based methods to gauge data utility, building on prior work (Shen, Raji, and Chen 2024). KL divergence (Kullback and Leibler 1951), also called *information gain* (Quinlan 1986), describes a measure of how much a model probability distribution Q is different from a true probability distribution P:

$$\mathrm{KL}(P||Q) = \int_{x \in \mathcal{X}} \log \frac{P(\mathrm{d}x)}{Q(\mathrm{d}x)} P(\mathrm{d}x) \qquad \text{(KL-Divergence)}$$

Because computing KL-divergence on hospital datasets $\mathcal{D}_o$ and $\mathcal{D}_i$ is non-trivial due to the high dimensionality of the data, Shen, Raji, and Chen (2024) proposes two groups of scores to make this divergence approximation tractable from small samples. We adopt one such score, $\mathrm{KL}_{\mathcal{XY}}$, which trains a classifier to distinguish between source and target samples and uses the classifier's output probabilities to estimate dataset divergence. We describe the $\mathrm{KL}_{\mathcal{XY}}$ score in detail in Section 4.2.

Privacy Model We operate under a semi-honest privacy model—also known as *honest-but-curious* or *passive security*—where parties follow protocols but may probe intermediate values. Parties are "curious", meaning that they can probe into the intermediate values to avoid paying for the data. This assumes a weaker security model than malicious security where a corrupted party may input foul data, but ensures the algorithm to be private throughout the computation. This privacy preservation model incentivizes collaboration, improving upon Shen, Raji, and Chen (2024) and Ilyas et al. (2022).

Dataset Divergence and Downstream Utility Building on the finding of Miller et al. (2021) where a model's in-distribution performance is related to its out-of-distribution performance (across all model choices), we confer that dataset divergence does predict downstream model's performance after combining the data (Table 3), using metrics introduced in Shen, Raji, and Chen (2024) (Section 4.2). Intuitively, in-distribution quality is paramount in low-data settings, where dataset divergence can capture greater complexity and nuance than accessing a few traits.Yet, privacy is unresolved: divergence measures entail accessing both entities' data (Shen, Raji, and Chen 2024; Ilyas et al. 2022), posing significant risks for heavily-regulated entities who are liable for any data exposure (Centers for Medicare & Medicaid Services 1996; Gervasi et al. 2022).

Secure Multiparty Computation (MPC) To cryptographically secure our divergence computation, we use Secure Multiparty Computation (MPC) (Yao 1982; Shamir 1979). MPC lets two or more parties to compute a function over private inputs, revealing only the final output (Goldwasser and Micali 2019).

Engineering machine learning workflows in private faces differs drastically from non-private machine learning engineering, due to 1. precision configurations and normalizations (to avoid blow-ups and underflows), 2. error control and performance engineering for machine learning in bit-limited spaces, and 3. debugging, hyperparameter tuning. Nevertheless, they are *provably* secure.

Despite their non-trivial engineering, MPC programs enjoy strong security guarantees and relative ease of deployment. Even small organizations can deploy MPC without any specialized hardware. Thus, the algorithms developed and shared in $\mathrm{SecureKL}$ readily enable zero-leakage dataset-to-dataset evaluations before sharing data.

Assumptions Consider high stakes domains where disparate data may have additive benefits to the existing data. To make privacy boundaries tractable, we make the following assumptions:

1. **Existing knowledge** is not private. The hospitals are aware of each other having such data to begin with. The

hospitals may know of the available underlying dataset size and format, which is assumed to be uniform across the hospitals in the setup to simulate unit-cost. Hospitals frequently know of each other's resources, and the available ICU units are contentious, not kept secret.

2. **Uniformity** of $|\mathcal{D}_i|$. Though each hospital gets to price their data and set their own budget, for generality, the uniformity assumption allows us to use the number of additional data sources n as the main "budget proxy" across different strategies.
3. **Legal risks** of sharing *any* data are omnipresent in high stakes domains. The risks with sharing sensitive data in data-leaking strategies, which we coin as π_d (demographic distance) and π_s (small sample), are not made explicit, but assumed to be "moderate" and "high" respectively. This abstraction side-steps legal discussion, which would go beyond the scope of our paper.
4. **No malice** is assumed on any of the parties involved, as each hospital wants to authentically sell their data and set up a potential collaboration. This assumption becomes stronger when the number of parties grows or when the setup changes to potentially more competitive industries with less trust. We note our limitations in Section 6.3.

4 Methods

	Strategy	Sub-strategies	Leakage
π_0	Blind (baseline)	n/a	zero
π_p	Private (SecureKL)	n/a	minimal
π_d	Demographic	sex, gender, race	moderate
π_s	Subset Sampling	1%, 10%, 100%	high

Table 1: **Concrete selection strategies π, differentiated by leakage risks**. For a strategy p, a set of targets is chosen, i.e., $T \leftarrow \pi(\mathcal{D}_o, n)$.

4.1 Defining Dataset Acquisition Strategies

Given the private and non-private data acquisition strategies illustrated in Figures 1 and 3, we associate privacy leakage as a primary cost of data partnerships. This section formalizes them by their leakage risks, summarized in Table 1.

A, high leakage, sharing raw data. $\pi_s(n, k)$ supposes each hospital to share a dataset of size k; a default setting of 1% is commonplace practice in some contracts, as a pre-requisite to being considered (Centers for Medicare & Medicaid Services 2025). Though leakage can be controlled through k, the data is inherently sensitive. The underlying distance metric follows the $\mathrm{KL}_{\mathcal{X}\mathcal{Y}}$ score introduced by Shen, Raji, and Chen (2024) , which we define in the following subsection.

B, moderate leakage, sharing summary statistics. $\pi_d(n)$ uses demographic metadata to guide data selection. This is implemented through ratio distance between source and target distributions, which may be considered aggregates therefore potentially not sensitive, such as when the underlying aggregation function ϕ is differentially private.

C, zero/minimal leakage, sharing no *additional* information besides what is assumed public, and what our method outputs, such as a score or a ranking. There are two methods: a. **Blind selection baseline**: $\pi_0(n)$ randomly selects n disjoint data sources, until data purchasing budget runs out. Prior works suggests that when $n = 1$, randomly selecting a source in hospital ICU may be risky and inefficient. b. **Our method** $\pi_p(n)$ selects data sources based on privacy-preserving measure for data combination, specifically Private $\mathrm{KL}_{\mathcal{X}\mathcal{Y}}$.

4.2 KL-based Methods, Without Privacy

In subset-sampling strategy $\pi_s(n, k)$, each of the candidate entities will leak a set of raw data. π_s is implemented with KL-based measures similar to Shen, Raji, and Chen (2024). Recall $\mathrm{KL}(P||Q)$ is not symmetrical, meaning that it is not a "metric" that satisfies triangle inequality. Intuitively, this means the measure is directional: a hospital's distribution P_o may be "close" to the target distribution P_i, but not the other way around:

$$\mathrm{KL}(P_o||P_i) = \int_{x \in \mathcal{X}} \log \frac{P_o(\mathrm{d}x)}{P_i(\mathrm{d}x)} P_i(\mathrm{d}x) \tag{Ideal Estimator}$$

Because we only have access to finite data $\mathcal{D}_o$ and $\mathcal{D}_i$, approximations are needed. Typically, a learned model can capture distributional information, used to estimate continuous entropy. Thus the joint distribution of features and labels from both the source and target are included, with the goal of deriving an efficient estimator for $\mathrm{KL}(P_o||P_i)$ that captures distributional shift from source to target.

Specifically, $\mathrm{KL}_{\mathcal{X}\mathcal{Y}}$ score used in $\mathrm{SecureKL}_{\mathcal{X}\mathcal{Y}}$ first trains a logistic regression model (Cox 1958) on $\mathcal{D}_o \cup \mathcal{D}_i$ – where the labels are folded into the covariates. The goal is to infer dataset membership using "proxy labels", defined as follows:

$$\mathcal{I}(x, y) = \begin{cases} 1 & \text{if } (x, y) \in \mathcal{D}_o \\ 0 & \text{if } (x, y) \in \mathcal{D}_i \end{cases}$$

A binary predictor is fit on this combined dataset, predicting $\mathcal{I}$ from $(\mathcal{X}, \mathcal{Y})$ using logistic regression. The model's output $p(x, y)$, also called the probability score, is $\mathrm{Score}(x, y)$.

A score of 0.5 or less means the datasets are not distinguishable, making the data potentially useful. Shen, Raji, and Chen (2024) established the insight that in data-limited domains of heterogeneous data sources, domain shifts of the covariates are useful for predicting whether the additional data helps the original task, similar to Miller et al. (2021). We note again that even though this model is trained on both parties' data, the final algorithm that the hospital uses to train on combined data is not restricted.

Then, the resulting model's probability score function $\mathrm{Score}(\cdot)$: $\mathcal{X}, \mathcal{Y} \to [0, 1]$ is averaged over a dataset in H_o, obtaining

$$\mathrm{KL}_{\mathcal{X}\mathcal{Y}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_o}(\mathrm{Score}(x, y)). \tag{KL-XY}$$

We reproduce Shen, Raji, and Chen (2024) 's results that $\mathrm{KL}_{\mathcal{X}\mathcal{Y}}$ is predictive of downstream change in AUC in hospital setting (Table 4). Let the score function g_{KL} be the

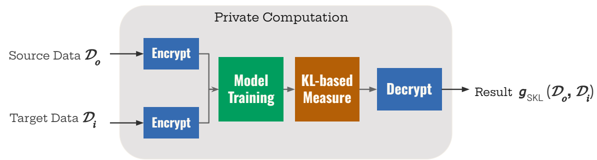

Figure 4: **Our method** SecureKL. Each side encrypts their data. Then, a model is *privately* trained on their joint data. Afterwards, their divergence is computed. Finally, only the final result of this dataset-to-dataset evaluation is revealed.

approximate of $\mathrm{KL}(\mathcal{D}_o||\mathcal{D}_i)$. The strategy selects the most likely hospital with the closest distance under the measure:

$$\pi_s(n=1, k=K) = \underset{i\in[1..N]}{\arg\min}\; g_{\mathrm{KL}}(\mathcal{D}_o, \mathcal{D}_i).$$

(KL-based Strategy, *in plaintext*)

When only a subset is available, this function is adjusted by swapping $\mathcal{D}_i$ for $\mathcal{D}'_i \subseteq \mathcal{D}_i$ where $|\mathcal{D}'_i| = k$. We denote the full dataset size as $K = |\mathcal{D}_i|$.

Though leakage can be controlled through k, yet the data is inherently sensitive. In ICU data, simulate that a default of 1% is shared, so $k = 3000 \times 1\% = 30$, though we run experiments with $k \in \{3, 30, 300, 3000\}$.

4.3 SecureKL: Private KL-based Method

Using MPC, we extend on $\mathrm{KL}_{\mathcal{XY}}$ to require no information sharing (besides what was already assumed public). Specifically we leverage private tensors and secure gradient descent in CrypTen (Knott et al. 2021) to implement private $\mathrm{KL}_{\mathcal{XY}}$. As illustrated in Figure 4, the logistic regression as well as the scoring need to be implemented in private.

Denote the private encoding of x as $[x]$.

$$\mathrm{SecureKL}_{\mathcal{XY}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_o}(\mathrm{Score}([x, y])).$$

(Secure KL-XY)

Let the score function g_{SKL} be the secure approximation of $\mathrm{KL}(\mathcal{D}_o||\mathcal{D}_i)$. The strategy selects the most likely hospital with the closest distance under the measure:

$$\pi_p(n=1) = \underset{i\in[1..N]}{\arg\min}\; g_{\mathrm{SKL}}(\mathcal{D}_o, \mathcal{D}_i).$$

(SecureKL Strategy, *encrypted*)

As shown in Figure 4, any KL-based measure g_{SKL} can be adapted to our setup. We mainly use $\mathrm{SecureKL}_{\mathcal{XY}}$ as the underlying measure. Its performance is detailed in Section 6.2. Additionally, even though our implementation measures distance of data between one source and one target party, the setup readily extends to accommodating multiple parties. Section 7.2 discusses potential deployment challenges.

4.4 Trivially "Private" Baseline: Blind Selection

$\pi_0(n)$ randomly selects n disjoint data sources, until data purchasing budget runs out. This random strategy may evade selection biases and help gather diverse data. Yet, prior work (Shen, Raji, and Chen 2024) suggests that $\pi_0(1)$ – randomly selecting one source – for ICU is risky and inefficient for mortality prediction.

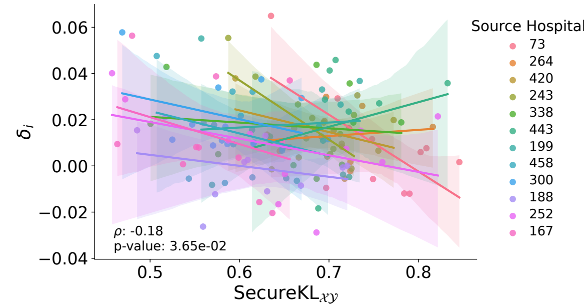

Figure 5: SecureKL**: Overall Correctness**. Rank correlation between SecureKL output and ground truth AUC change, δ_i, from acquiring 1 additional dataset for a given source hospital H_o. We propose selecting data partner ranked by our secure system under $\mathrm{SecureKL}_{\mathcal{XY}}$ score to reliably increase AUC gains. ($|\mathbf{H}| = 12$ hospitals; colored by source.)

4.5 Alternative Data-leaking Baseline: Sharing Summary Statistics

A relaxation to sharing no sensitive data is to share metadata. While demographic traits are often *causal* and available, their exact cause in relation to the task is not a priori established (without a highly effective model), therefore their success in distributional-matching is not guaranteed to be strong. Additionally, in practice, the most effective model that results from data combination may or may not be causally-sound. Nevertheless, we posit alternative strategy $\pi_d(n)$ to find the demographically close candidates to guide data selection: Let $\phi : \mathcal{D} \to \mathbb{R}^m$ be an m-dimensional summary statistic of a demographic feature i.e. the racial distribution of patients. Then, we use the distributional distance between $\mathcal{D}_o$ and $\mathcal{D}_i$, characterized by their L_2-distance through ϕ:

$$\pi_d(n=1) = \underset{i\in[1..N]}{\arg\min}\; L_2(\phi(\mathcal{D}_o)||\phi(\mathcal{D}_i)).$$

(Demographic-based Strategy)

5 Experiment Setup

5.1 Experimental Questions

1. **Consistency**: Does MPC degrade the original measure's effectiveness? Since MPC implementations introduce approximations, we first validate SKL's correctness by examining the correlation of **private scores** (SecureKL) vs. **plaintext scores** ($\mathrm{KL}_{\mathcal{XY}}$, with full data access). Then, for the consistency of choosing encryption in the hospital domains, we examine plaintext and encrypted versions' correlation with downstream ground truth rankings across hospitals.
2. **Positivity**: Does our method pick entities that reliably improve performance? If source dataset D_o can only add data from n more sources, does our measure lead to eventual AUC improvements? Specifically, in multi-dataset combination, we examine whether using SKL can improve the source hospital's downstream outcome. When

selecting a single (or a few) additional data source, how many hospitals improve with our method? Additionally, we compare our method with alternative, privacy-leaking strategies.

3. **Error analysis**: If our privacy-preserving method is not the dominant strategy against alternatives including limited data accessibility, why? As we know, small and uncertain improvements for downstream tasks underscore the inherent difficulty of evaluating data utility without seeing the full data. We are especially interested in analyzing (a) hospitals with low $\mathrm{SecureKL}_{\mathcal{XY}}$ and $\mathrm{KL}_{\mathcal{XY}}$ correlations, and (b) hospitals lagging AUC improvements using the random strategy π_0 or limited-sample strategy π_s.

5.2 MPC implementation

$\mathrm{SecureKL}_{\mathcal{XY}}$ includes training logistic regression model in private. We implement measures based on dataset divergence by building a custom logistic regression model over encrypted data leveraging CrypTen. The model parameters and input are encoded as 16-bit MPCTensors, ensuring that all computations, including forward passes, sigmoid activations, and gradient descent updates, are performed in private.

Additional baseline details We additionally run our experiments on plaintext methods used in Shen, Raji, and Chen (2024), including the $\mathrm{KL}_{\mathcal{X}}$ measure, which is similar to $\mathrm{KL}_{\mathcal{XY}}$ without using each data source's labels:

$$\mathrm{KL}_{\mathcal{X}} = \mathbb{E}_{(x)\sim\mathcal{D}_o}(\mathrm{Score}(x)). \quad \text{(KL-X)}$$

To compare against $\mathrm{KL}_{\mathcal{X}}$ for baseline, we additionally implemented its encrypted version, $\mathrm{SecureKL}_{\mathcal{X}}$, as a g_{SKL} candidate.

$$\mathrm{SecureKL}_{\mathcal{X}} = \mathbb{E}_{(x)\sim\mathcal{D}_o}(\mathrm{Score}([x])). \quad \text{(Secure KL-X)}$$

Optimizers Because L-BFGS – the optimizer prior work (Shen, Raji, and Chen 2024) used in plaintext-only with Scikit-Learn (Pedregosa et al. 2011) – is not available as an encrypted version, our MPC experiments are facilitated with SGD optimizer. A fair comparison between the scores obtained through plaintext and encrypted settings necessitates re-implementing plaintext scores, $\mathrm{Score}(X)$ and $\mathrm{Score}(X, Y)$, using logistic regression with SGD in PyTorch (Paszke et al. 2019). The hyper-parameter tuning for SGD in private and plain text are performed independently, as they do not transfer. Appendix C will discuss hyperparameter-tuning in detail.

5.3 Data and Model Setup

eICU dataset. The downstream task is 24-hour mortality prediction from ICU data using the eICU Collaborative Research Dataset (Pollard et al. 2018). This dataset contains $> 200,000$ real-world admission records from 208 hospitals across the United States.

A note on data filtering. Because medical research is inherently complex, ensuring reproducibility on statistical methods can be significantly challenging. Water et al. (2023) proposes that the research community follow a shared set of tasks with fixed preprocessing pipelines that are clinically informed – essentially a machine learning training and evaluation protocol on eICU – in order to facilitate method verification on the same benchmark. Therefore, our work follows their data cleaning criteria and the evaluation protocol. Additionally, we use the hospital exclusion criteria in (Shen, Raji, and Chen 2024) to obtain top 12 hospitals, where the most patient visits are collected (each with > 2000 patients).

Downstream model and baselines for eICU. Recall that each strategy uses the same K number of records per hospital – in our experiment, $K = 3000$. For π_s which leaks a subset k of all samples, a default $k = 1\%|\mathcal{D}_i|$ randomly drawn samples are shared. In ICU data, we run experiments on $\{0.1\%, 1\%, 10\%, 100\%\}$.

Following holistic benchmarking tools in (Water et al. 2023), our strategy comparisons take 1500 samples and the downstream model performance – AUC_o, AUC_T – uses 400 samples (unless otherwise noted). Specifically, the AUC change, δ_i or δ_T, comes from 1. combining 1500 random samples from each selected dataset and 2. combining it with 1500 samples from $\mathcal{D}_o$, and 3. subtracting the baseline model's AUC[1].

Folktables dataset Though we primarily focus on hospital domain, we additionally validate using Folktables (Ding et al. 2021), predicting across 15 states an individual's annual income exceeds \$50,000[2]. The details of our processing, which diverges from that of eICU, is included in Appendix F.

6 Results and Analysis

We organize our results around three research questions: whether using multiparty implementation sacrifices $\mathrm{KL}_{\mathcal{XY}}$'s efficacy (**consistency**; Section 6.1), whether our method reliably picks hospitals that improve performance (**positivity**; Section 6.2), and where our method may fail (**error analysis and limitations**; Section 6.3).

6.1 Consistency Between Plaintext and Encrypted Computations

Because our encrypted computation is the first implementation of dataset divergence in MPC, we ought to show that $\mathrm{SecureKL}_{\mathcal{XY}}$ and $\mathrm{SecureKL}_{\mathcal{X}}$ lead to highly comparable behaviors as $\mathrm{KL}_{\mathcal{XY}}$ and $\mathrm{KL}_{\mathcal{X}}$.

Spearman's Rank Correlation Coefficient for Underlying Scores For each source hospital H_o, use all full samples for $\mathcal{D}_i$. Between $\mathrm{KL}_{\mathcal{XY}}$ and $\mathrm{SecureKL}_{\mathcal{XY}}$ on $\mathcal{D}_o$ and $\mathcal{D}_i$ for all remaining hospitals H_i, $\mathbb{E}_{H_o\sim\mathbf{H}}[\rho] = 0.908$ with a range of $[0.691, 1.0]$, obtaining $p < 0.02$ across all hospitals. Between $\mathrm{SecureKL}_{\mathcal{X}}$ and $\mathrm{KL}_{\mathcal{X}}$, $\mathbb{E}_{H_o\sim\mathbf{H}}[\rho] = 0.9303$ with a range of $[0.455, 0.991]$, with 11 of 12 hospitals achieving p-values below 0.05. After applying Hochberg false discovery rate correction (Benjamini and Hochberg 1995), our

[1]The samples are fixed across all experiments, the sample numbers are chosen to match (Shen, Raji, and Chen 2024)'s setup.

[2]Because the Folktables dataset contains simpler features and is low-dimensional, KL divergence is directly computed rather than estimated. For implementation details, see Appendix F.

p-values remain significant. This range may be an artifact of sweeping hyperparameters independently in plaintext and encrypted optimisations, because sharing the same SGD hyperparameters would result in a tighter range. For all 12 hospitals, see appended Appendix D for details.

Impact of Adding Security on AUC Correlation We further examine the effect by *adding* encryption through its impact on the downstream AUC, using how AUC improvements are ranked. This rank is compared with how secure measures (i.e., $\mathrm{SecureKL}_{\mathcal{XY}}$) and plaintext measures (i.e. $\mathrm{KL}_{\mathcal{XY}}$) rank hospitals. This comparison investigates the extent of the shift in the full hospital ranking, when we switch from a plaintext setup to encrypted. For $H_o \sim \mathbf{H}$, we measure δ_i that results from adding $\mathcal{D}_i$ to $\mathcal{D}_o$ for all i. This correlates all target hospitals $\{H_i\}$ with their ground truths $\{\delta_i\}$ in the case of picking a single target hospital. We find the linear coefficient for encrypted $\mathrm{SecureKL}_{\mathcal{XY}}$ to be -0.182 and plaintext $\mathrm{KL}_{\mathcal{XY}}$ to be -0.184 (99% matching). Both $\mathrm{SecureKL}_{\mathcal{X}}$ and $\mathrm{KL}_{\mathcal{X}}$ have a linear coefficient of -0.164 with δ_i. For all strategies' correlations with ground truth at $n = 1$, see Appendix B.

6.2 Positivity in Realistic Setup

We apply $\mathrm{SecureKL}$ in a pragmatic multi-source data combination problem, where each strategy acquires n datasets for $n \in \{1, 2, 3\}$.

Overall Positivity. For $n = 1$, we find that π_p improves AUC in 10 out of the 12 hospitals. When $n = 2$ and $n = 3$, we find that using π_p consistently improves AUC for all hospitals. Overall, 34 out of the 36 dataset combinations we evaluate on have an AUC improvement $\delta_T > 0$, suggesting that π_p is a reasonable strategy for selecting hospital dataset combinations with a high expected return $\mathbb{E}[P_{H_o \sim \mathbf{H}}(\delta_T > 0)]$ for the source hospital from using our strategy.

Comparing With Alternative Strategies Other strategies – π_0, π_d, and π_s – can also arrive at "positive" datasets. Comparing private method to other strategies at $n = 3$, i.e, $\pi_p(n = 3)$, we describe our results in Figure 6:

1. π_p (our method based on $\mathrm{SecureKL}_{\mathcal{XY}}$) has a median δ_T of 0.020, and a standard deviation of 0.015. Our results indicate that for 50% of the hospitals, π_p gives a $\delta_T >= .02$. Compared to other strategies, π_p has the highest median, the lowest standard deviation, and it is one of two strategies that improves performance for all hospitals.
2. Demographic-based strategies underperform compared to π_p on average. However, we observe that π_d-gender can be highly effective for a subset of hospitals, as it achieves the highest 75th percentile (Q3) of 0.033 among all strategies. This indicates that for 25% of hospitals, $\delta_T \geq 0.033$. Despite this, π_d-gender has a lower median value of 0.012 compared to π_p, exhibits a high standard deviation (0.022), and degrades the performance for certain hospitals. Similarly, π_d-age has a median of 0.014, and π_d-race has a median of 0.008, both lower than π_p's median.
3. Plaintext small-sample strategies, π_s, outperform all demographic-based methods but slightly underperform relative to π_p. For instance, $\pi_s(k = 300)$ has a median δ_T of 0.0178, and although it achieves $\delta_T > 0$ across all hospitals, it performs worse on average compared to π_p and exhibits a higher standard deviation (0.017). $\pi_s(k = 30)$ has a median δ_T of 0.0165. Compared to other strategies, it has the largest standard deviation (0.024), and it degrades the performance for some hospitals.

In summary, our method π_p achieves the highest AUC improvement on average with the lowest standard deviation, demonstrating **more consistent improvements** for all hospitals. While the average improvement of π_p is small, demographic-based and plaintext small-sample strategies exhibit greater variability, with some strategies improving performance for specific subsets of hospitals but underperforming or degrading results in others.

6.3 SecureKL Error Analysis

Underlying Data Limitations In high-stakes domains, data partnerships are expensive, but potentially detrimental – this forms a challenging landscape for evaluating methods on real-world data. Indeed, as shown in Table 2, the AUC gain is small across all strategies, and the variance is high. This suggests that 3 hospitals' data is likely still too small for the general task to the robust explains limited AUC gains, highlighting the need to maximize samples for informative decisions. The key distinction, however, is that privacy-leaking methods (demographic, small sample) and blind baseline risk performance declines in many hospitals while SKL consistently improves downstream tasks **more reliably** than alternatives, across all hospitals, over multiple runs.

Underlying Score Limitations Data addition algorithms underpin the effectiveness of our method. Even if $\mathcal{D}_o$ obtains access to all the plaintext data, there is no guarantee that π_p can correctly predict whether the data is useful. As seen in Figure 8, Hospital 243's utility when acquiring another data set is badly correlated with plaintext and encrypted KL-XY scores. This leads to its bad strategy for acquiring the top 3 hospitals, as seen in the middle pane of Figure 9. Interestingly, for this hospital, no other informational strategy excels, either, so choosing a random 3 may be preferred.

This behavior stems from the underlying measure, not from adding secure computation: in Figure 8 and Figure 9, the encrypted performance closely follows that of plaintext performance, for both good and bad downstream correlations.

Sometimes, Not All Underlying Data Is Needed Relatedly, when seeing a few samples can successfully identify useful candidate hospitals, π_s (which is on small samples) outperforms π_p (which is on full samples).

In the right panel of Figure 9, hospital 199, the smaller sample sizes achieve a score that better reflects ground truth as a data addition strategy. In that case, the hospital may not need the full sample to know which target hospitals to collaborate with.

Dataset	n	Demographic Summary Strategies			Subset-sampling Strategies		Secure Strategy (**Ours**)
		π_d-gender	π_d-age	π_d-race	$\pi_s(k=300)$	$\pi_s(k=30)$	π_p SecureKL$_{\mathcal{XY}}$
eICU	1	**0.020** ± 0.023	0.012 ± 0.016	0.014 ± 0.015	0.012 ± 0.014	0.010 ± 0.017	0.011 ± 0.018
	2	0.016 ± 0.016	0.017 ± 0.017	0.015 ± 0.013	0.024 ± 0.020	0.017 ± 0.019	**0.027** ± 0.022
	3	0.020 ± 0.023	0.016 ± 0.019	0.011 ± 0.021	0.021 ± 0.017	0.021 ± 0.024	**0.024** ± 0.015
Folktables	1	0.006 ± 0.017	0.007 ± 0.013	0.006 ± 0.009	0.006 ± 0.016	0.007 ± 0.016	**0.009** ± 0.013
	2	**0.015** ± 0.019	**0.015** ± 0.014	0.010 ± 0.013	0.011 ± 0.014	0.014 ± 0.018	0.010 ± 0.017
	3	0.013 ± 0.018	0.014 ± 0.016	0.013 ± 0.017	0.014 ± 0.016	**0.016** ± 0.019	0.011 ± 0.017

Table 2: **AUC improvements in mean and standard deviation**, across all source regions for each strategy π, for eICU and Folktables setups. n denotes the number of candidate datasets added to the source dataset. The small gains and high variance from adding selected datasets highlight the precarious nature of assessing data value in the real world. **Bold** indicates highest AUC improvement per n. *Note*: Only π_p SecureKL$_{\mathcal{XY}}$ is private.

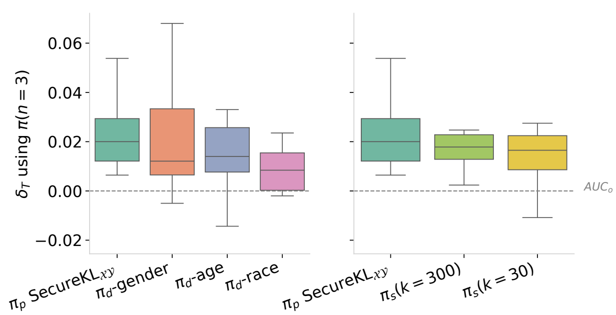

Figure 6: **AUC change δ_T over all strategies in eICU prediction** (higher is better). Our private dataset evaluation strategy π_p outperforms demographic-based strategy π_d **(left)**, and subset-sampling strategy π_s for $k = 300$ (10%) and $k = 30$ (1%) **(right)**, after combining source data with the top 3 candidates.

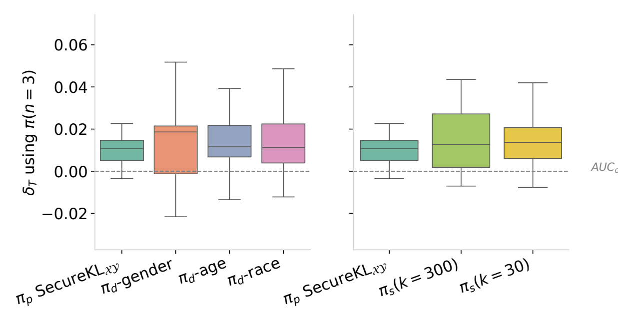

Figure 7: **AUC change δ_T over all strategies in Folktables dataset prediction** (higher is better). All strategies exhibit comparable distributions, after combining data from top 3 candidates. In a noisy domain, our method is stable: it neither excels nor penalizes against non-private strategies.

This behavior is specific to the interaction of the data and the underlying score, and does not affect the general insight that adding private computation preserves privacy (and eases privacy-related risks that hinder data sharing). We further note that our method still clearly applies to encrypted computation on a smaller data set under data minimization.

7 Discussion

7.1 SecureKL Contribution

After establishing that π_p with SecureKL$_{\mathcal{XY}}$is a robust strategy in practical downstream performance, we hereby summarize the benefits of SecureKL and elaborate on their practical implications.

Matching Plaintext Performance in Downstream Tasks. Our major contribution is to match plaintext performance with no data sharing. Using MPC provides *input privacy*, meaning that if both hospitals only want to know the resulting score, the computation can be done without leaking original data. This strong guarantee can significantly ease the tension related to privacy and compliance in setting up a collaboration, leading to a practical "data appraisal stage" in data-limited high stakes domains.

Gain from Data Availability. In contrast to limited-sample approaches, a key advantage for our method π_p is that it takes advantage of all the underlying data – generally impossible with non-secure methods for private data in heavily regulated domains. The general intuition is that data is localized; therefore, once a good target hospital is identified, we would prefer to acquire all of the data. It may be tempting to assert that we prefer the highest k for data addition algorithms as well. In our experiments, while this is generally true, the smaller k sometimes outperform larger k in plaintext strategy π_s, which we investigate in Section 6.3 and in Figure 9. This occasionally non-monotonic behavior mirrors the challenge of data combination itself: even within one source dataset for the same estimator, more data is not necessarily better. This suggests the potential for a hospital-specific alternative to sharing a large amount of data for some source hospital, and points to future directions to using secure computation on a minimal-sized sample dataset for minimal performance overhead while remaining private.

Potential Improvements to SecureKL In the case where that output can be sensitive, i.e., when a source hospital queries a target hospital multiple times and accrues information through the score function, the *output* can also be

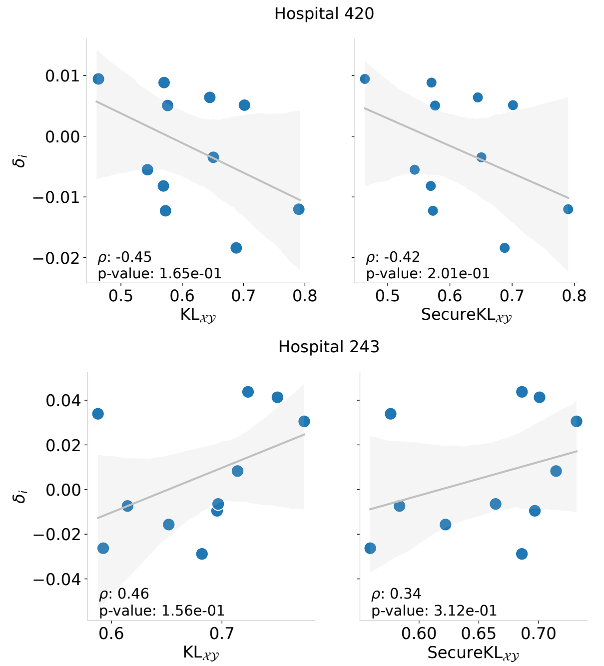

Figure 8: **The correlation between AUC change δ_i and underlying score $\mathrm{KL}_{\mathcal{XY}}$ affects $\mathrm{SecureKL}_{\mathcal{XY}}$'s efficacy.** In Hospital 420, the $\mathrm{KL}_{\mathcal{XY}}$ scores identify beneficial data ($\rho < 0$). For Hospital 243, $\mathrm{KL}_{\mathcal{XY}}$ anti-correlates ($\rho > 0$).

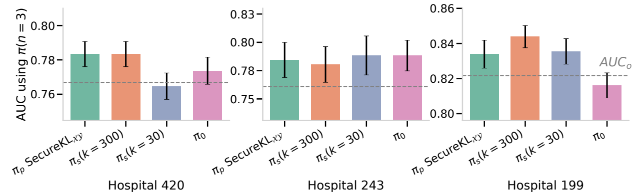

Figure 9: Mean AUC ($\pm$ standard deviation). **Left**: $\mathrm{SecureKL}_{\mathcal{XY}}$ outperforms $\pi_s(k=30)$ and π_0. **Middle**: All strategies perform similarly. **Right**: $\pi_s(k=300)$ outperforms $\mathrm{SecureKL}_{\mathcal{XY}}$.

made privacy-preserving through differentially private data releases, such as using randomized response (Dwork, Roth et al. 2014).

7.2 Potential Challenges to Broader Adoption

Our code is readily usable by small organizations. While our approach generalizes to model-based measures (by substituting g_{SKL}) and scales to multiple parties, our work also uncovered deployment limitations.

1. **Operational**: engineering personnel limitations. While our implementation requires little cryptographic knowledge to deploy, it still needs technically-trained staff at each participating hospital to collaborate and maintain. This skill is similar to using pre-packaged software, cleaning data, and setting up network calls.
2. **Engineering Extensions**: Extending any MPC protocol is non-trivial, as security engineering is a specialized skill. While SecureKL applies broadly to other underlying scores in multi-party setups, *validating* a new MPC algorithm requires software engineering – prototyping, tuning, debugging – and numerical verification – akin to data analytics and research - likely requiring technical talents who can be especially costly for hospitals to retain in-house.
3. **Framework Limitation**: While CrypTen is designed to accommodate PyTorch, it is a research tool where not all plain text functionalities are implemented. For example, writing optimizers – such as L-BFGS – and custom operators that are not readily available requires both machine learning and cryptography knowledge. Moreover, the protocol incurs additional computational overhead, especially if hyper-parameters become more complex to sweep[3]. This will likely improve with time, as new frameworks address design shortcomings.
4. **Inherent to Secure Computation**: When the method requires significant hyper-parameter tuning, such as using SGD on small batch data with learning rate schedules, plaintext tuning may not transfer perfectly. As detailed in Appendix C, our hyperparameters for SGD differ in encrypted and plaintext settings. Thus, as encrypted computation *hides* loss curves and training details by default, development is expected to be complex. This is because both hospitals want to ensure model fit with secure evaluation, but may not want to expend the computational cost of private hyperparameter sweeping.

8 Related Works

We briefly relate alternative approaches towards dataset privacy for sharing. More work on data pricing, differential privacy, and private fine-tuning are included in Appendix G.2.

Augmenting existing data with synthetic data in medical domains Synthetic data generation has emerged as a promising approach to expand training datasets while preserving privacy. Generative adversarial networks (GANs) have shown success in generating realistic cancer incidence data (Goncalves et al. 2020), medical imaging data (Thambawita et al. 2022), and electronic health records (Baowaly et al. 2019). These methods preserve statistical properties of the original data while providing differential privacy guarantees. Transforming data into a similar form that desensitizes certain attributes can be desirable (Drechsler 2011; Howe et al. 2017; Nikolenko 2021; Gonzales, Guruswamy, and Smith 2023; Sweeney 2002). Yet, to still preserve the utility of the dataset transformed for analytics or learning tasks is challenging by itself (Jordon et al. 2021). Additionally, outside the scope of sensitive data that is transformed, little privacy guarantee is available, leading to re-identification risks (Narayanan and Shmatikov 2006; Jordon et al. 2021).

In addition, evaluation of synthetic medical data reveals challenges in capturing rare conditions and maintain-

[3]For our work, the performance metrics are provided in Appendix E for reference

ing consistent relationships between multiple health variables (Goncalves et al. 2020). For tabular data, methods like CTGAN and TVAE (Xu et al. 2019) have demonstrated ability to learn complex distributions while preserving correlations between features. However, these approaches often struggle with high-dimensional data and can introduce subtle biases that impact downstream model performance (Assefa et al. 2020). Recent work has also explored combining synthetic data with differential privacy to provide formal privacy guarantees (Jordon, Yoon, and Van Der Schaar 2018). While these methods offer stronger privacy protection, they often face significant utility loss, particularly for rare but important cases in the original dataset (Yang et al. 2024a).

Secure Data Combination Recent work has explored methods for securely combining datasets while preserving privacy and improving model performance. Early approaches focused on using secure multi-party computation to enable multiple parties to jointly train models without sharing raw data (Aono et al. 2017). However, these methods often struggled with computational overhead and communication costs when dealing with large-scale datasets (McMahan et al. 2017). More recent techniques have introduced frameworks for evaluating potential data partnerships before commitment. These approaches use privacy-preserving protocols to estimate the compatibility and complementarity of different datasets (Leung, Law, and Sima 2019; Chakraborty et al. 2024). Some methods focus specifically on measuring distribution shifts between datasets without revealing sensitive information (Duan et al. 2021). Others trained the downstream model in private, but limited to LASSO (van Egmond et al. 2021). Several systems have been developed to facilitate secure data combination in specific domains. In healthcare, methods have been proposed for securely combining patient records across institutions while maintaining HIPAA compliance (Raisaro et al. 2018; van Egmond et al. 2021). Financial institutions have explored similar approaches for combining transaction data while preserving client confidentiality (Liu et al. 2021).

Federated Learning Cross-silo federated, decentralized, and collaborative ML (McMahan et al. 2017; Li et al. 2020; Bonawitz et al. 2019; Kairouz et al. 2021) focus on acquiring more data through improved data governance and efficient system design. Healthcare machine learning is considered especially suitable, as health records are often isolated (Rieke et al. 2020; Xu et al. 2021; Nguyen et al. 2022; Cho et al. 2025). Yet, even though no raw data is shared, model parameters or gradients flow through the system. As the federated computing paradigm offers no privacy guarantee, the system is vulnerable to model inversion (Geiping et al. 2020) and gradients leakage attacks (Boenisch et al. 2023; Zhu, Liu, and Han 2019). A subtle but urgent concern is that privacy risks discourage the very formation of the federation when optimization is traded off with privacy (Lyu et al. 2022; Raynal and Troncoso 2024). Building on the insight that useful data is often disparately-owned, we tackle the specific incentive problem between pairs of players where one side trains the model, instead of scaling up the *number* of parties through a federation.

Compared to vanilla Federated Learning, an MPC system (Shamir 1979; Yao 1982; Bonawitz et al. 2017; Knott et al. 2021) provides stronger guarantee in terms of input security. Model owners and data owners can potentially federate their proprietary data, including model weights, training, and testing data, can work together under stringent privacy requirements. Our work extends the line of work by (Xu, Hannun, and Van Der Maaten 2022; Yang et al. 2024b; Bonawitz et al. 2017) that demonstrates the potential of incorporating MPC in various federated scenarios. On the practical side, unlike mobile-based networks for secure federated learning protocols (Bonawitz et al. 2019), our system assumes a smaller number of participants, where communication cost and runtime are not dominant concerns.

Comparing with Federated Learning with Privacy Guarantees Privacy incentivizes federation (Usynin, Rueckert, and Kaissis 2024). Specifically, preserving privacy between the parties under federated learning uses secure computation methods (Truex et al. 2019; Bonawitz et al. 2017; Stripelis et al. 2021; Cho et al. 2025; Froelicher et al. 2021) combined with differential privacy (Xu et al. 2023; Usynin, Rueckert, and Kaissis 2024; Avent et al. 2017), trusted execution environments (Mo et al. 2021) – an approach that is coined "Privacy-in-Depth" in Kairouz et al. (2021).

Performing dataset evaluation with MPC, as in SecureKL, can be seen as an extension of data federation, by adding a separate privacy-preserving component *before* all parties commit to an entire federated learning system. It is low-commitment, as privacy is preserved by default, and there is no requirement to continue in any system; it tackles the incentives' problem blocking collaboration. By demonstrating utility, the two parties foster trust. Our work complements the line of ambitious systems that incorporate MPC in potential federated scenarios where the participants are semi-honest (Yang et al. 2024b; Bonawitz et al. 2017; Cho et al. 2025).

9 Conclusion

Our work demonstrates that privacy-preserving data valuation can help organizations identify beneficial data partnerships while maintaining data sovereignty. Through SecureKL, we show that entities can make informed decisions about data sharing without compromising privacy or requiring complete dataset access. As the AI community continues to grapple with data access challenges, particularly in regulated domains like healthcare, methods that balance privacy and utility will become increasingly critical for responsible advancement of the field. As noted in Section 6.3, our approach has several limitations, including the fact that, despite impressive aggregate results, our method is less effective for individual hospitals, which motivates future work. Our method assumes static datasets and may not generalize well to scenarios where data distributions evolve rapidly over time. A sequential version of our framework may more closely model dynamic data collaborations. Future work should explore extending these techniques to handle more complex data types and dynamic distribution shifts while maintaining strong privacy guarantees.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Acemoglu, D.; Makhdoumi, A.; Malekian, A.; and Ozdaglar, A. 2022. Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, 14(4): 218–256.

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; and McDermott, M. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78.

Aono, Y.; Hayashi, T.; Wang, L.; Moriai, S.; et al. 2017. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE transactions on information forensics and security*, 13(5): 1333–1345.

Assefa, S. A.; Dervovic, D.; Mahfouz, M.; Tillman, R. E.; Reddy, P.; and Veloso, M. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, 1–8.

Avent, B.; Korolova, A.; Zeber, D.; Hovden, T.; and Livshits, B. 2017. {BLENDER}: Enabling local search with a hybrid differential privacy model. In *26th USENIX Security Symposium (USENIX Security 17)*, 747–764.

Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; and Shmatikov, V. 2020. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, 2938–2948. PMLR.

Baowaly, M. K.; Lin, C.-C.; Liu, C.-L.; and Chen, K.-T. 2019. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3): 228–241.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological*, 57: 289–300.

Boenisch, F.; Dziedzic, A.; Schuster, R.; Shamsabadi, A. S.; Shumailov, I.; and Papernot, N. 2023. When the curious abandon honesty: Federated learning is not private. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 175–199. IEEE.

Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konecnỳ, J.; Mazzocchi, S.; McMahan, H. B.; et al. 2019. Towards Federated Learning At Scale: System Design. *arXiv preprint arXiv:1902.01046*.

Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191.

Bradley, V. C.; Kuriwaki, S.; Isakov, M.; Sejdinovic, D.; Meng, X.-L.; and Flaxman, S. 2021. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890): 695–700.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Centers for Medicare & Medicaid Services. 1996. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/.

Centers for Medicare & Medicaid Services. 2025. Limited Data Set (LDS) Files — CMS.

Chakraborty, N.; Sharma, A.; Dutta, J.; and Kumar, H. D. 2024. Privacy-Preserving Data Quality Assessment for Time-Series IoT Sensors. In *2024 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 51–57. IEEE.

Cheon, J. H.; Kim, A.; Kim, M.; and Song, Y. 2017. Homomorphic encryption for arithmetic of approximate numbers. In *Advances in Cryptology–ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I 23*, 409–437. Springer.

Cho, H.; Froelicher, D.; Chen, J.; Edupalli, M.; Pyrgelis, A.; Troncoso-Pastoriza, J. R.; Hubaux, J.-P.; and Berger, B. 2025. Secure and federated genome-wide association studies for biobank-scale datasets. *Nature Genetics*, 1–6.

Compton, R.; Zhang, L.; Puli, A.; and Ranganath, R. 2023. When More is Less: Incorporating Additional Datasets Can Hurt Performance By Introducing Spurious Correlations. In Deshpande, K.; Fiterau, M.; Joshi, S.; Lipton, Z.; Ranganath, R.; Urteaga, I.; and Yeung, S., eds., *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, 110–127. PMLR.

Cook, R. D. 1979. Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365): 169–174.

Cox, D. R. 1958. The Regression Analysis of Binary Sequences. *Journal of the royal statistical society series b-methodological*, 20: 215–232.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Neural Information Processing Systems*.

Drechsler, J. 2011. *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media.

Duan, M.; Liu, D.; Ji, X.; Wu, Y.; Liang, L.; Chen, X.; Tan, Y.; and Ren, A. 2021. Flexible clustered federated learning for client-level data distribution shift. *IEEE Transactions on Parallel and Distributed Systems*, 33(11): 2661–2674.

Dwork, C. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, 1–12. Springer.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Feldman, V.; and Zhang, C. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33: 2881–2891.

Froelicher, D.; Troncoso-Pastoriza, J. R.; Raisaro, J. L.; Cuendet, M. A.; Sousa, J. S.; Cho, H.; Berger, B.; Fellay, J.; and Hubaux, J.-P. 2021. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nature communications*, 12(1): 5910.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33: 16937–16947.

Gervasi, S. S.; Chen, I. Y.; Smith-McLallen, A.; Sontag, D.; Obermeyer, Z.; Vennera, M.; and Chawla, R. 2022. The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It. *Health Affairs*, 41(2): 212–218.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, 2242–2251. PMLR.

Goldwasser, S.; and Micali, S. 2019. Probabilistic encryption & how to play mental poker keeping secret all partial information. In *Providing sound foundations for cryptography: on the work of Shafi Goldwasser and Silvio Micali*, 173–201.

Goncalves, A.; Ray, P.; Soper, B.; Stevens, J.; Coyle, L.; and Sales, A. P. 2020. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20: 1–40.

Gonzales, A.; Guruswamy, G.; and Smith, S. R. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1): e0000082.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.

Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G. F.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; and Zhou, Y. 2017. Deep Learning Scaling is Predictable, Empirically. *ArXiv*, abs/1712.00409.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 30016–30030.

Howard, J.; and Ruder, S. 2018. Universal language model fine-tuning for text classification. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, 328–339. Association for Computational Linguistics.

Howe, B.; Stoyanovich, J.; Ping, H.; Herman, B.; and Gee, M. 2017. Synthetic data for social good. *arXiv preprint arXiv:1710.08874*.

Huang, L.; Dou, Y.; Liu, Y.; Wang, J.; Chen, G.; Zhang, X.; and Wang, R. 2021. Toward a research framework to conceptualize data as a factor of production: The data marketplace perspective. *Fundamental Research*, 1(5): 586–594.

Ilyas, A.; Park, S. M.; Engstrom, L.; Leclerc, G.; and Madry, A. 2022. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1167–1176. PMLR.

Jordon, J.; Jarrett, D.; Saveliev, E.; Yoon, J.; Elbers, P.; Thoral, P.; Ercole, A.; Zhang, C.; Belgrave, D.; and van der Schaar, M. 2021. Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification. In *NeurIPS 2020 Competition and Demonstration Track*, 206–215. PMLR.

Jordon, J.; Yoon, J.; and Van Der Schaar, M. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *ArXiv*, abs/2001.08361.

Knott, B.; Venkataraman, S.; Hannun, A.; Sengupta, S.; Ibrahim, M.; and van der Maaten, L. 2021. Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34: 4961–4973.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, 5637–5664. PMLR.

Koutris, P.; Upadhyaya, P.; Balazinska, M.; Howe, B.; and Suciu, D. 2015. Query-based data pricing. *Journal of the ACM (JACM)*, 62(5): 1–44.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Kumar, A.; Finley, B.; Braud, T.; Tarkoma, S.; and Hui, P. 2020. Marketplace for AI models. *arXiv preprint arXiv:2003.01593*.

Le Scao, T.; Wang, T.; Hesslow, D.; Bekman, S.; Bari, M. S.; Biderman, S.; Elsahar, H.; Muennighoff, N.; Phang, J.; Press, O.; et al. 2022. What Language Model to Train if You Have One Million GPU Hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 765–782.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Leung, C.; Law, A.; and Sima, O. 2019. Towards privacy-preserving collaborative gradient boosted decision trees. *UC Berkeley*.

Li, J.; Fang, A.; Smyrnis, G.; Ivgi, M.; Jordan, M.; Gadre, S. Y.; Bansal, H.; Guha, E.; Keh, S. S.; Arora, K.; et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37: 14200–14282.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

Liang, F.; Yu, W.; An, D.; Yang, Q.; Fu, X.; and Zhao, W. 2018. A survey on big data market: Pricing, trading and protection. *Ieee Access*, 6: 15132–15154.

Liu, Z.; Huang, D.; Huang, K.; Li, Z.; and Zhao, J. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, 4513–4519.

Lyu, L.; Yu, H.; Ma, X.; Chen, C.; Sun, L.; Zhao, J.; Yang, Q.; and Philip, S. Y. 2022. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*.

L'Heureux, A.; Grolinger, K.; Elyamany, H. F.; and Capretz, M. A. M. 2017. Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, 5: 7776–7797.

Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and Van Der Maaten, L. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, 181–196.

McKinney, S. M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G. S.; Darzi, A.; et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788): 89–94.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Meng, X.-L. 2018. Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2): 685–726.

Miller, J. P.; Taori, R.; Raghunathan, A.; Sagawa, S.; Koh, P. W.; Shankar, V.; Liang, P.; Carmon, Y.; and Schmidt, L. 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, 7721–7735. PMLR.

Mo, F.; Haddadi, H.; Katevas, K.; Marin, E.; Perino, D.; and Kourtellis, N. 2021. PPFL: Privacy-preserving federated learning with trusted execution environments. In *Proceedings of the 19th annual international conference on mobile systems, applications, and services*, 94–108.

Narayanan, A.; and Shmatikov, V. 2006. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.

Nguyen, D. C.; Pham, Q.-V.; Pathirana, P. N.; Ding, M.; Seneviratne, A.; Lin, Z.; Dobre, O.; and Hwang, W.-J. 2022. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)*, 55(3): 1–37.

Nguyen, T.; Van Nguyen, C.; Lai, V. D.; Man, H.; Ngo, N. T.; Dernoncourt, F.; Rossi, R. A.; and Nguyen, T. H. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Nikolenko, S. I. 2021. *Synthetic data for deep learning*, volume 174. Springer.

Papernot, N.; Thakurta, A.; Song, S.; Chien, S.; and Erlingsson, Ú. 2021. Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9312–9321.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*, 2613.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Petty, J.; van Steenkiste, S.; and Linzen, T. 2024. How Does Code Pretraining Affect Language Model Task Performance? *arXiv preprint arXiv:2409.04556*.

Pollard, T. J.; Johnson, A. E. W.; Raffa, J. D.; Celi, L. A.; Mark, R. G.; and Badawi, O. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning*, 1: 81–106.

Radić, M.; Herrmann, P.; Stein, T.; Heirich, N.; and Radić, D. 2024. Pricing Data Based on Value: A Systematic Literature Review. In *Proceedings of the Future Technologies Conference*, 319–339. Springer.

Raisaro, J. L.; Troncoso-Pastoriza, J. R.; Misbach, M.; Sousa, J. S.; Pradervand, S.; Missiaglia, E.; Michielin, O.; Ford, B.; and Hubaux, J.-P. 2018. M ed C o: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4): 1328–1341.

Raynal, M.; and Troncoso, C. 2024. On the Conflict of Robustness and Learning in Collaborative Machine Learning. *arXiv preprint arXiv:2402.13700*.

Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S. G.; Novikov, A.; Barth-Maron, G.; Gimenez, M.; Sulsky, Y.; Kay, J.; Springenberg, J. T.; et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H. R.; Albarqouni, S.; Bakas, S.; Galtier, M. N.; Landman, B. A.; Maier-Hein, K.; et al. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1): 1–7.

Shamir, A. 1979. How to share a secret. *Communications of the ACM*, 22(11): 612–613.

Sheller, M. J.; Edwards, B.; Reina, G. A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R. R.; et al. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1): 12598.

Shen, J. H.; Raji, I. D.; and Chen, I. Y. 2024. The Data Addition Dilemma. In *Machine Learning for Healthcare Conference*. PMLR.

Shen, Z.; Tao, T.; Ma, L.; Neiswanger, W.; Liu, Z.; Wang, H.; Tan, B.; Hestness, J.; Vassilieva, N.; Soboleva, D.; et al. 2023. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*.

Stripelis, D.; Saleem, H.; Ghai, T.; Dhinagar, N.; Gupta, U.; Anastasiou, C.; Ver Steeg, G.; Ravi, S.; Naveed, M.; Thompson, P. M.; et al. 2021. Secure neuroimaging analysis using federated learning with homomorphic encryption. In *17th international symposium on medical information processing and analysis*, volume 12088, 351–359. SPIE.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05): 557–570.

Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599.

Tenopir, C.; Allard, S.; Douglass, K.; Aydinoglu, A. U.; Wu, L.; Read, E.; Manoff, M.; and Frame, M. 2011. Data sharing by scientists: practices and perceptions. *PloS one*, 6(6): e21101.

Thambawita, V.; Salehi, P.; Sheshkal, S. A.; Hicks, S. A.; Hammer, H. L.; Parasa, S.; Lange, T. d.; Halvorsen, P.; and Riegler, M. A. 2022. SinGAN-Seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5): e0267976.

Tramer, F.; and Boneh, D. 2020. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*.

Truex, S.; Baracaldo, N.; Anwar, A.; Steinke, T.; Ludwig, H.; Zhang, R.; and Zhou, Y. 2019. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security*, 1–11.

Usynin, D.; Rueckert, D.; and Kaissis, G. 2024. Incentivising the federation: gradient-based metrics for data selection and valuation in private decentralised training. In *European Interdisciplinary Cybersecurity Conference*, 179–185.

van Egmond, M. B.; Spini, G.; van der Galiën, O.; IJpma, A.; Veugen, T.; Kraaij, W.; Sangers, A.; Rooijakkers, T.; Langenkamp, P.; Kamphorst, B. A.; van de L'Isle, N.; and Kooij-Janic, M. 2021. Privacy-preserving dataset combination and Lasso regression for healthcare predictions. *BMC Medical Informatics and Decision Making*, 21.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Philip, S. Y. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8): 8052–8072.

Water, R. V. D.; Schmidt, H.; Elbers, P.; Thoral, P. J.; Arnrich, B.; and Rockenschaub, P. 2023. Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML. *ArXiv*, abs/2306.05109.

Xu, J.; Glicksberg, B. S.; Su, C.; Walker, P.; Bian, J.; and Wang, F. 2021. Federated learning for healthcare informatics. *Journal of healthcare informatics research*, 5: 1–19.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32.

Xu, X.; Hannun, A.; and Van Der Maaten, L. 2022. Data appraisal without data sharing. In *International Conference on Artificial Intelligence and Statistics*, 11422–11437. PMLR.

Xu, Z.; Zhang, Y.; Andrew, G.; Choquette-Choo, C. A.; Kairouz, P.; McMahan, H. B.; Rosenstock, J.; and Zhang, Y. 2023. Federated learning of gboard language models with differential privacy. *arXiv preprint arXiv:2305.18465*.

Yang, M.; Chi, C.-H.; Lam, K.-Y.; Feng, J.; Guo, T.; and Ni, W. 2024a. Tabular Data Synthesis with Differential Privacy: A Survey. *arXiv preprint arXiv:2411.03351*.

Yang, Z.; Zhang, Y.; Zheng, Y.; Tian, X.; Peng, H.; Liu, T.; and Han, B. 2024b. FedFed: Feature distillation against data heterogeneity in federated learning. *Advances in Neural Information Processing Systems*, 36.

Yao, A. C. 1982. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, 160–164. IEEE.

Zhang, Y.; Qin, J.; Park, D. S.; Han, W.; Chiu, C.-C.; Pang, R.; Le, Q. V.; and Wu, Y. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.