Fold-stratified cross-validation for unbiased and privacy-preserving federated learning

Romain Bey, Romain Goussault, François Grolleau, Mehdi Benchoufi, Raphaël Porcher

STRATIFIED CROSS-VALIDATION FOR UNBIASED AND PRIVACY-PRESERVING FEDERATED LEARNING

A PREPRINT

Romain Bey
Université de Paris
CRESS, INSERM, INRA
F-75004 Paris, France

Romain Goussault
Service Oncodermatologie
CHU Nantes, CIC 1413, CRCINA Inserm 1232
Nantes, France

Mehdi Benchoufi
Université de Paris
CRESS, INSERM, INRA
F-75004 Paris, France

Raphaël Porcher
Université de Paris
CRESS, INSERM, INRA
F-75004 Paris, France
raphael.porcher@aphp.fr

ABSTRACT

Objective

We introduce *stratified cross-validation*, a validation methodology that is compatible with privacy-preserving federated learning and that prevents data leakage caused by duplicates of electronic health records (EHR).

Materials and methods

Stratified cross-validation complements cross-validation with an initial stratification of EHR in folds containing similar patients, thus ensuring that duplicates of a record are jointly present either in training or in validation folds. Monte Carlo simulations are performed to investigate the properties of *stratified cross-validation* in the case of a model data analysis.

Results

In situations where duplicated EHR could induce over-optimistic estimations of accuracy, applying *stratified cross-validation* prevented this bias, while not requiring full deduplication. However, a pessimistic bias might appear if the covariate used for the stratification was strongly associated with the outcome.

Discussion

Although *stratified cross-validation* presents low computational overhead, to be efficient it requires the preliminary identification of a covariate that is both shared by duplicated records and weakly associated with the outcome. When available, the hash of a personal identifier or a patient's date of birth provides such a covariate. On the contrary, pseudonymization interferes with *stratificatied cross-validation* as it may break the equality of the stratifying covariate among duplicates.

Conclusion

Stratified cross-validation is an easy-to-implement methodology that prevents data leakage when a model is trained on distributed EHR that contain duplicates, while preserving privacy.

***K*eywords** Federated Learning · Privacy · Validation · Duplicated Electronic Health Records · Data Leakage

Background and significance

The large-scale collection of data and its analysis by artificial intelligence (AI) algorithms have led to new scientific discoveries and huge expectations for the near future [1–5]. Although AI algorithms (random forest, gradient boosting, neural networks, etc. [6]) provide powerful tools, they are difficult to develop as they generally require large training datasets to reach reasonable performances [7] and often detailed, high dimensional records about individuals. Beyond the technical challenges that such a data collection represents, storing and analyzing a large amount of personally identifying information (PII) may moreover imply serious risks regarding privacy, and recent research projects have been hindered by public opinion concerns [8, 9]. These concerns are not unfounded as re-identification attacks have regularly broken the anonymity of large datasets [10–13]. To address these risks new regulations that impose higher security standards have been introduced [14]. Technically, it moreover appears necessary to complement classical anonymization techniques as they are intrinsically limited in the case of high dimensional data [15, 16]. Handling and analyzing securely such data therefore requires moving from the *anonymize, release and forget* approach to configurations where a data curator *secures and controls* the use of data that remain to some extent identifying [17]. In the latter configuration, the question arises as to which organization should play the role of the trusted data curator. In the case of medical records, patients and IT managers have been reluctant to devote this role to centralized private or public organisations [8, 9, 18], limiting large-scale research on electronic health records (EHR). To solve this issue a technique called federated learning has recently been proposed. This technique enables the training of AI models while keeping records in decentralized trusted data warehouses curated for instance by hospitals [19–25]. Federated learning appears as a promising privacy-enhancing technique that avoids single points of failure, and it is currently being developed and tested in various projects worldwide [26–30].

In addition to privacy concerns, recent controversies indicate that many AI models may have been validated improperly, shedding doubt on the performances that have been advertized [31–35]. One of the most frequent sources of bias in performance estimation is the data leakage that occurs when data used for validation and training are correlated [36–39]. Avoiding data leakage requires building training and validation datasets in such a way that all the data related to a given individual are exclusively in the former or the latter. This dataset building procedure may be compromised by the presence of different records related to the same individual: this risk is far from being negligible as it has been shown that up to 15% of records in medical information systems are duplicates [40] and that many patients have records in multiple hospitals [41]. Data leakage induced by duplicates may be especially important in the case of AI models, as they are often trained on large real-world datasets such as EHR, that have not been curated for research [4]. To address data leakage caused by duplicates, deduplication algorithms, often called record linkage algorithms, have been developed that rely on various deterministic or probabilistic methods [42].

Although federated learning and deduplication algorithms address privacy and validation issues respectively, they cannot be easily combined. Indeed, deduplication relies on the comparison of PII through the computation of a similarity index established between two potentially duplicated records, whereas federated learning avoids PII exchange and therefore prevents their comparison when records are located in different data warehouses. Consequently, detecting duplicates of a given record that are present in two different hospitals while preserving their privacy appears challenging. Some protocols have been developed and deployed to enable privacy-preserving deduplication in a federated learning setting, but they still require further research to be scalable [43–45]. In this paper, we propose an easier-to-implement approach that makes it possible to avoid data leakage in a federated learning setting while not relying on deduplication algorithms. We consider the classical cross-validation technique for performance estimation, and complement it by an initial stratification of datasets. Whereas some stratification techniques have already been combined with cross-validation in order to limit disparities between randomly chosen folds [46], in this article we extend the use of stratification to avoid data leakage in the case of undetected data duplicates.

This article is organised as follows. In *Materials and Methods*, we describe the *stratified cross-validation* methodology we propose and the synthetic datasets we study. In *Results*, we simulate a data analysis in a federated learning setting following different validation strategies, the performances and limitations of which are detailed in *Discussion* section.

Materials and methods

Stratified cross-validation

We consider a model f that computes a predicted $\tilde{y}$ of an outcome y using covariates $\boldsymbol{x} = (x_1, x_2, ..., x_m)$: $\tilde{y} = f(\boldsymbol{x})$. We consider a performance metric that we want to maximize and that is computed as the expectation of a function $h(y, \tilde{y})$. In the case of accuracy $h(y, \tilde{y}) = I(y = \tilde{y})$, where the function $I(A)$ equals 1 if A is true and 0 otherwise. A record $\boldsymbol{r} = (\boldsymbol{x}, y)$ is a point in a mathematical space Ω that gathers for an individual her covariates $\boldsymbol{x}$ and her realized outcome y, and a dataset $\mathcal{D}$ is a collection of records. A dataset $\mathcal{D}$ contains duplicates when there are two records $\boldsymbol{r}_i$

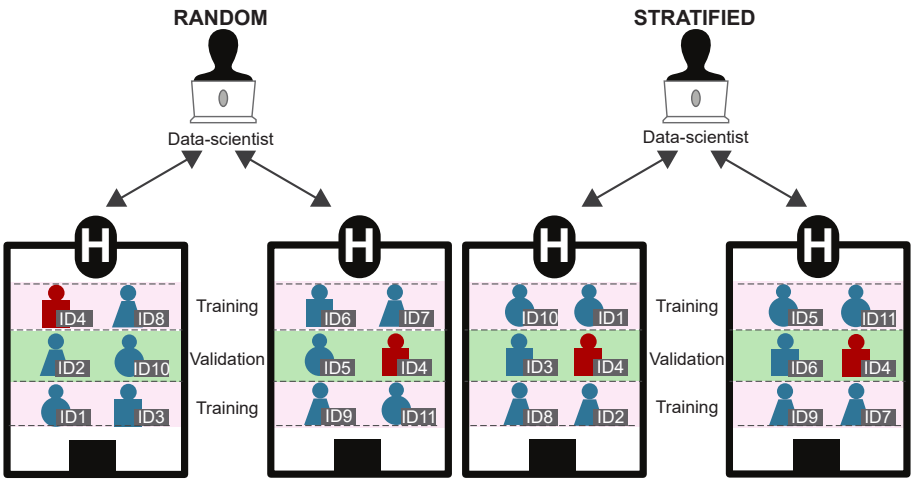

Figure 1: Privacy-preserving federated learning: analysis by a data scientist of medical records (blue and red individuals) distributed in two hospitals without extracting personally identifying information (PII). One individual's record is duplicated in the two hospitals (red, ID4), due for instance to multiple admissions. The performances of a model are estimated through cross-validation, partitioning the datasets in training and validation folds either randomly (left) or through stratification, *i.e.* grouping similar patients in folds (right). Whereas duplicated records (red) may be simultaneously in training and validation folds when random partitioning is applied, thus causing data leakage, this risk is circumvented by stratification.

and $\boldsymbol{r}_j$ with $i \neq j$ such that $\boldsymbol{r}_i = \boldsymbol{r}_j$. For the sake of simplicity we limit ourselves to exact duplicates but as discussed later our conclusions apply also to inaccurate duplicates, caused for instance by flawed or incomplete recording of data related to a given individual. We consider hereafter that there are never two individuals with perfectly identical records implying that equal records always relate to the same individual that has been registered more than once. We moreover consider that individuals whose records are duplicated are distributed uniformly in the population.

Cross-validation is a statistical technique commonly used to estimate the performances of a model [6]. In cross-validation a dataset $\mathcal{D}$ is partitioned in k folds (disjoint subsets) $\mathcal{D}_i$, with $i = 1, .., k$. For each fold $\mathcal{D}_i$ the statistical model is trained on all the folds apart from fold $\mathcal{D}_i$ (training), and its performances are estimated on fold $\mathcal{D}_i$ (validation). The average of per-fold estimates is used as the estimate of the performances of the same model that would have been trained on all the records. Classical cross-validation often relies on a random partitioning of the dataset in k folds. In that case duplicated records may be simultaneously present in training and validation folds thus inducing data leakage and yielding over-optimistic estimates of performance compared to cross-validation without duplicates, that we consider hereafter as the unbiased estimate. Although in that case data leakage is a consequence of duplicated records, it is possible to avoid it without complete deduplication. Indeed, data leakage occurs only when duplicated records are used both for training and validation. Ensuring that all duplicated records related to a given individual are used only for training or only for validation therefore prevents the risk of data leakage.

We consider training and validation of a model in a federated learning setting where records are distributed in different hospitals. In that case, many datasets related to different hospitals are jointly analysed by an external data scientist without exchanging PII (see Figure 1). Cross-validation in a federated learning setting involves partitioning the dataset $\mathcal{D}^{(\alpha)}$ of each hospital α in k folds $\mathcal{D}_i^{(\alpha)}$ with $i = 1, .., k$. Folds are then merged over hospitals to obtain global folds: $\mathcal{D}_i = \cup_\alpha \mathcal{D}_i^{(\alpha)}$. To better characterize the presence of duplicates in a federated learning setting we introduce the following definitions:

Definition 1. Intra-hospital deduplication: For all records $\boldsymbol{r}_i, \boldsymbol{r}_j \in \mathcal{D}^{(\alpha)}$ in a hospital α with $i \neq j$ we have $\boldsymbol{r}_i \neq \boldsymbol{r}_j$.

Definition 2. Inter-hospital deduplication: For all records $\boldsymbol{r}_i \in \mathcal{D}^{(\alpha)}$, $\boldsymbol{r}_j \in \mathcal{D}^{(\beta)}$ in different hospitals $\alpha \neq \beta$ we have $\boldsymbol{r}_i \neq \boldsymbol{r}_j$.

Datasets $\mathcal{D}$ that jointly fulfill definitions 1 and 2 are completely deduplicated: *i.e.* there are no two identical records $\boldsymbol{r}_i = \boldsymbol{r}_j$ with $i \neq j$ in the whole dataset. As explained above, although many deduplication techniques have been developed to fulfill definition 1, fulfilling definition 2 remains challenging without loosing the privacy-enhancing

advantage of federated learning. We therefore consider a weaker definition of deduplication that is sufficient to avoid data leakage between folds:

Definition 3. Inter-fold deduplication: For all records $\boldsymbol{r}_i \in \mathcal{D}_m^{(\alpha)}$, $\boldsymbol{r}_j \in \mathcal{D}_n^{(\beta)}$ related to different fold indexes $m \neq n$ we have $\boldsymbol{r}_i \neq \boldsymbol{r}_j$.

Definition 3 is weaker than definitions 1 and 2 as it can be fulfilled without removing all duplicates if one ensures instead that duplicates of a given record are present in the same fold. We propose hereafter a technique to create folds that fulfill definition 3. We consider a partition of the record space Ω in k subspaces Ω_i: $\Omega = \cup_{i=1}^{k} \Omega_i$ and $\Omega_i \cap \Omega_j = \emptyset$ if $i \neq j$. Such a partition can be realized stratifying Ω relatively to one covariate, and we refer to such a partition as a stratification. Once a stratification has been defined, each hospital α dataset $\mathcal{D}^{(\alpha)}$ can be partitioned in folds $i = 1, 2, ..., k$ as follows:

$$\begin{cases} \mathcal{D}_1^{(\alpha)} = & \mathcal{D}^{(\alpha)} \cap \Omega_1 \\ & ... \\ \mathcal{D}_i^{(\alpha)} = & \mathcal{D}^{(\alpha)} \cap \Omega_i \\ & ... \\ \mathcal{D}_k^{(\alpha)} = & \mathcal{D}^{(\alpha)} \cap \Omega_k \end{cases} \tag{1}$$

Partitioning each hospital dataset $\mathcal{D}^{(\alpha)}$ using a given stratification leads for $i \neq j$ to: $\mathcal{D}_i^{(\alpha)} \cap \mathcal{D}_j^{(\beta)} = (\mathcal{D}^{(\alpha)} \cap \Omega_i) \cap (\mathcal{D}^{(\beta)} \cap \Omega_j) = (\mathcal{D}^{(\alpha)} \cap \mathcal{D}^{(\beta)}) \cap (\Omega_i \cap \Omega_j) = \emptyset$ and the definition 3 is therefore fulfilled. Combining stratification technique equation (1) with classical cross-validation constitutes the validation methodology that we call *stratified cross-validation*.

Although *stratified cross-validation* prevents over-optimistic estimations of performance induced by duplicates, it does not systematically provide an unbiased estimator compared to cross-validation in the absence of duplicates. A stratification procedure may indeed induce training and validation folds featuring different covariate distributions and covariate-outcome associations. A model trained and validated on such folds tends to overfit the training population and to be unfit for a generalization to the validation population, and *stratified cross-validation* appears akin to the external validation on a new population. Validating externally a model on a population coming from a different hospital is commonly recognized as a proof of quality as it measures the generalizability of a model to new care contexts, but *stratified cross-validation* unfortunately does not measure this relevant inter-hospital generalizability as folds cannot be identified with hospitals. The stratification procedure should therefore be defined as to minimize the irrelevant pessimistic bias associated with the heterogeneity of folds populations. An ideal stratifying covariate would therefore be a covariate shared by duplicates but fully independent of the other covariates and of the outcome, as it would provide folds that would be statistically equivalent. Such a stratifying covariate is often not available as one only records covariates that are associated to some extent with patient's medical condition, and a challenge of *stratified cross-validation* consists in finding a surrogate stratifying covariate that is weakly correlated to the other covariates and to the outcome. In the following section we run simulations to investigate and discuss the impact of various stratification strategies.

Stratified cross-validation

- Choose or create a stratifying covariate x_{str} that is weakly associated with the other covariates and with the outcome.
- Define thresholds t_0, t_1, ..., t_k with k the number of folds, such that there are approximately the same amount of records fulfilling $t_i < x_{str} < t_{i+1}$ for each i.
- Associate each record with the fold index i that fulfills $t_i < x_{str} < t_{i+1}$.
- Group all records with the same fold index i in inter-hospital folds $\mathcal{D}_i$ and apply cross-validation on these folds.

Simulation

In order to study the properties of *stratified cross-validation* we simulate data analysis in a federated learning setting in presence of duplicates. We generate synthetic datasets in which a binary outcome y depends on 10 covariates

$x_1, x_2, ..., x_{10}$. Covariates are generated randomly following a multivariate gaussian distribution:

$$\begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_{10} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_{10} \end{bmatrix}, \Sigma \right) \tag{2}$$

with $\mu_1, \mu_2, ..., \mu_{10}$ the covariates means and Σ the covariance matrix. To generate Σ we choose 10 eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_{10})$, and sample a random orthogonal matrix O of size 9×9. An intermediate covariance matrix Σ' is obtained through:

$$\Sigma' = O \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \lambda_9 \end{bmatrix} O^t \tag{3}$$

The final covariance matrix Σ is then generated concatenating Σ' with λ_{10} in a block-diagonal matrix:

$$\Sigma = \left[\begin{array}{c|c} \Sigma' & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline \begin{matrix} 0 \cdots 0 \end{matrix} & \lambda_{10} \end{array}\right] \tag{4}$$

Generating covariates $\boldsymbol{x} = (x_1, x_2, ..., x_{10})$ according to equations (2), (3) and (4) provides a set of 9 correlated covariates $x_1, x_2, ..., x_9$ and an independent covariate x_{10}. Once covariates have been generated, we randomly generate their associated outcomes y through a logistic model. We consider a situation where the logarithm of the odds is a strongly non-linear function of the covariates:

$$\log \frac{p\big(y=1|x_1,...,x_{10}\big)}{p\big(y=0|x_1,...,x_{10}\big)} = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_1 x_2 + a_5 x_3 I(x_4 > 0) + a_6 x_5^2 I(x_6 > 0) + a_7 x_7 I(x_8 x_9 > 0) \tag{5}$$

with $(a_0, a_1, ..., a_7)$ a set of constants. Covariates x_1, x_2, ..., x_9 are associated with the outcome y contrary to x_{10} that remains independent of all other variables. The strongly non-linear case given by equation (5) corresponds to a generic situation with complex interactions that cannot be accounted for by simple generalized linear models.

Each simulation consists in generating randomly $n_{gen} = 10000$ records and then in adding randomly $n_{dup} = 2000$ duplicates (17% of the total number of records). Each of the n_{gen} original records are drawn from the probability distributions given by equations (2) and (5), and is then attributed randomly to one of the $n_h = 5$ hospitals with uniform probability. To generate duplicates we randomly draw one of the n_{gen} original records and one of the n_h hospitals. We then add a duplicate of the drawn record to the drawn hospital unless a duplicate of the original record already exists in the hospital that has been drawn, thus ensuring that definition 1 is fulfilled. We repeat this procedure until n_{dup} duplicated records have been added to the hospitals datasets.

Unless stated otherwise we consider centered covariates $(\mu_1, \mu_2, ..., \mu_{10}) = (0,\ 0, ...,\ 0)$ generated through equation (2) using eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_{10}) = (1.0,\ 1.2,\ 1.4,\ 1.6,\ 1.8,\ 2.0,\ 2.2,\ 2.4,\ 2.6,\ 2.8)$. Outcomes are generated using equation (5) with parameters $(a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7) = (-2,\ 0.4,\ 0.8,\ 1.2,\ 0.4,\ 1.2,\ 3.0,\ 2.0)$ leading to $\sim 47\%$ of records associated to a positive outcome $y = 1$. Cross-validation is realized with $k = 5$ folds. The code used for the simulations is available in the Supplemental material.

Results

Model definition and dataset partitioning strategies

High-dimensional non-linear problems on tabular data are commonly modeled using gradient boosting [6] and we use here its implementation in XGBoost library [47]. We consider trees of depth 3 that are added successively during 200 boosting iterations. The learning parameter is set to 0.6 and we use the binary logistic loss function. Although it is currently not possible to apply directly XGBoost in a federated learning setting, protocols are being developed to circumvent this difficulty [48, 49]. These computational considerations are not related to the data leakage issue under consideration and, for the sake of simplicity, in our simulations XGBoost is applied on physically centralized datasets simulating in this way gradient boosting in a federated learning setting.

For each simulation we first generate a dataset without duplicates and we fit and validate a gradient boosting model using the classical cross-validation methodology, measuring thus an unbiased estimate of performances. We then add

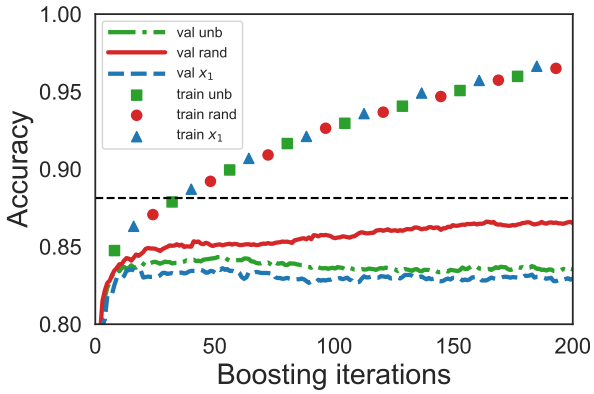

Figure 2: Accuracies computed through cross-validation as a function of the number of boosting iterations. Symbols and curves correspond respectively to training accuracies and validation accuracies. Green, red and blue colors correspond respectively to *unbiased*, *random* and *stratified along* x_1 fold-partitioning strategies. *Unbiased* validation accuracy lies between the over-optimistic *random* and the pessimistic x_1-*stratified* estimates. Horizontal black dashed line indicates the theoretical optimal accuracy $Accuracy_{opt}$.

duplicates and consider various validation strategies. *Random partitioning* consists in partitioning randomly each hospital dataset in k folds of the same size. *Stratified partitioning* consists in choosing first a stratifying covariate x_{str} and a set of thresholds $\{t_0, t_1, ..., t_k\}$. Each record is then attributed to the fold i that fulfills $t_i < x_{str} < t_{i+1}$. We choose thresholds t_i in such a way that the number of records in each global fold $\mathcal{D}_i$ is the same. Once fold-partitioning is realized, gradient boosting models are fitted and validated on these folds.

Model training and validation

For each $i \in 1, .., k$ the model is trained on all the folds apart from fold $\mathcal{D}_i$ and its training performances are measured on the same folds. Figure 2 shows the training learning curves obtained during boosting for *unbiased* (squares), *random* (circles) and *stratified along* x_1 (triangles) strategies. Training accuracies increase monotonously as the model learns from the training dataset, and the learning speed does not depend on the fold-partitioning strategy that is adopted. Indeed, training accuracies do not depend on strategy-dependent data leakage between training and validation folds.

For each $i \in 1, .., k$ the model trained on folds $\mathcal{D}_j$ with $j \neq i$ is validated on fold $\mathcal{D}_i$. Figure 2 shows the variation of validation accuracies during training (solid and dashed lines). When a deduplicated dataset is used (green dashed curve), the unbiased validation accuracy increases during the first 50 boosting iterations and then saturates at a value that is lower than the optimal accuracy that a predictive model could reach (straight dashed black line). As expected the validation accuracy remains lower than the optimal accuracy. When duplicates are added and *random* fold-partitioning strategy is used, the estimated validation accuracy (solid red curve) is biased by data leakage and it increases monotonously until falsely reaching a high accuracy. When *stratified along* x_1 strategy is adopted, duplicates of a given record are grouped in the same fold and definition 3 is fulfilled: there is consequently no over-optimistic bias and the estimated accuracy remains close to the unbiased one, but a small pessimistic bias is observed that is due to inter-fold heterogeneity. Applying *random* fold-partitioning strategy on a dataset with duplicates therefore misses the saturation of the performances observed in *unbiased* and *stratified* cases.

Bias and feature importance

We ran additional simulations to better understand the implications of the choice of a stratifying covariate. Figure 3 shows validation accuracies obtained after 200 boosting iterations for 30 simulations using the same set of generating parameters Σ and $(a_0, a_1, ..., a_7)$ but applying various fold-partitioning strategies. The distribution of estimated accuracies are shown as violin plots. Whereas *random* fold-partitioning always leads to over-optimistic estimates of accuracy (red) compared to the *unbiased* estimates obtained without duplicates (green), other stratification strategies lead to accuracy estimates that feature pessimistic biases of variable importance (blue). Whereas x_5 stratification leads to a pessimistic bias of roughly 5%, stratifying along x_1, x_2, x_4, x_8, x_9 or x_{10} leads to estimates that are close to the unbiased ones. The arbitrary choice of a stratifying covariate may consequently lead to important pessimistic biases. The closeness with the unbiased estimates of the estimates obtained having stratified along x_{10} was predictable as x_{10} is an independent variable corresponding thus to an ideal stratifying covariate that does not induce inter-fold heterogeneity.

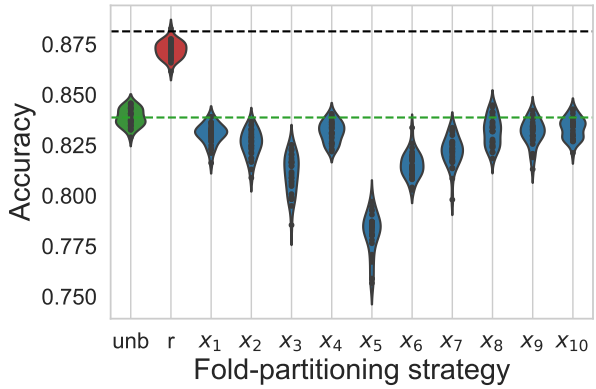

Figure 3: Violin plots for cross-validation estimates of accuracy adopting either an *unbiased* (green), a *random* (red) or a $x_1, x_2, ..., x_{10}$-*stratified* (blue) fold-partitioning strategy and running 30 simulations. Horizontal black and green dashed lines correspond respectively to the optimal accuracy that a model could reach $Accuracy_{opt}$ and to the mean unbiased estimate of the accuracy $Accuracy_{unb}$ reached by the model under consideration. Whereas *random* fold-partitioning leads to over-optimistic estimates of accuracy, $x_1, x_2, ..., x_{10}$-*stratified* estimates feature pessimistic biases of various sizes.

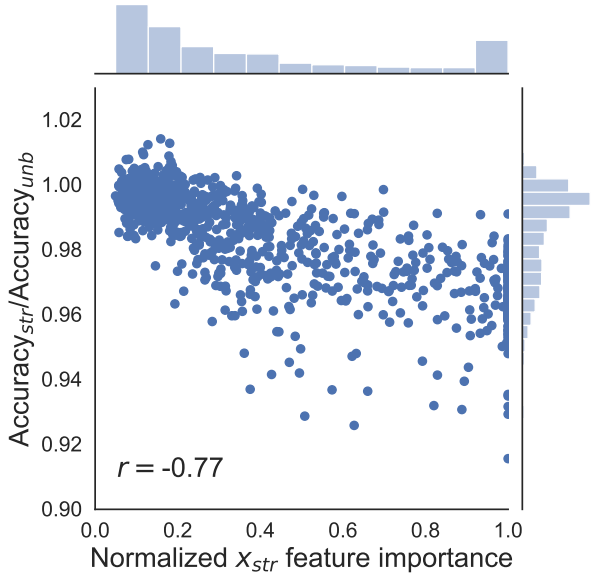

Figure 4: Ratio of x_{str}-*stratified* estimate of accuracy over the *unbiased* estimate of accuracy plotted with respect to the normalized importance of the stratifying covariate x_{str} (*see text*). 100 datasets are generated corresponding to different $\{\Sigma, \boldsymbol{a}\}$ and, for each dataset, each covariate is taken successively as stratifying covariate. A Pearson correlation coefficient $r = -0.77$ is measured.

To better characterize pessimistic biases we run additional simulations varying also the covariate covariance matrix Σ and the outcome generating parameters $\boldsymbol{a} = (a_0, a_1, ..., a_7)$. To generate each new covariance matrix Σ we draw random eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{10}$ from a uniform distribution on the interval $[1, 3]$, we generate a new random orthonormal matrix O and we combine them through equations (3) and 4. Each new outcome generating parameter a_i, $i = 0, .., 6$ is drawn from a uniform distribution on $[-5, 5]$. We draw 100 different sets of parameters $\{\Sigma, \boldsymbol{a}\}$ and, for each set of generating parameters, we generate a dataset. For each dataset, we fit and validate a gradient boosting model using classical cross-validation and measuring its unbiased validation accuracy $Accuracy_{unb}$. We then add duplicates and fit successively 10 gradient boosting models corresponding each to *stratified cross-validation* having stratified along one of the 10 covariates. Figure 4 shows for each one of the 1000 models the validation accuracy $Accuracy_{str}$ measured having stratified along x_{str} divided by the unbiased estimate $Accuracy_{unb}$ obtained on the same dataset, plotted with respect to normalized feature importance of x_{str} measured during the training of the unbiased model (cf. [6] for feature importance computation). We observe a negative Pearson correlation coefficient of -0.77 between the pessimistic bias on accuracy and the normalized importance of the stratifying covariate: the stronger the importance of the stratifying covariate, the bigger the pessimistic bias.

Discussion

The data leakage phenomenon at play in our simulations is specific neither to the XGBoost model nor to the synthetic data under scrutiny, and the risk of duplicate-caused data leakage should be addressed whatever the scientific problem at stake. *Stratified cross-validation* methodology avoids data leakage between folds by fulfilling inter-fold deduplication (definition 3) without requiring full deduplication (definitions 1 and 2), providing thus a validation methodology that is robust to the presence of undetected duplicates (Figures 2 and 3). Although *stratified cross-validation* avoids an over-optimistic bias due to data leakage between folds, it may be subject to a pessimistic bias due to inter-folds heterogeneity (Figures 2 and 3). In order to limit this undesired pessimistic bias it appears optimal to choose a stratifying covariate that is weakly associated to the other covariates and to the outcome (Figure 4). Determining *a priori* which covariate is weakly or strongly associated to the other covariates and to the outcome is challenging, and can rely either on prior knowledge of the problem under scrutiny or on preliminary fitting of the model on a local dataset. In the special case of the date of birth or a personal identifier being available, hashing it provides an ideal stratifying covariate that is shared by duplicates. Another difficulty might arise when records related to a given individual are not perfectly equal as it is the case when duplicated records correspond to different recording events such as hospital admissions. But perfect equality of duplicated records is not necessary to apply *stratified cross-validation*: identifying a single stratifying covariate, the value of which is shared among records related to a given individual, is indeed sufficient.

Stratified cross-validation methodology presents generic limitations. Firstly when duplicated records are so different that no shared covariate can be found to stratify, it appears impossible to ensure through stratification that two records related to a given individual are present in the same fold. Secondly when data curators are not hospitals but directly individuals, such as for instance in the case of data stored in mobile phones [25], it is impossible to partition datasets in folds as each dataset corresponds to a single record. Thirdly pseudonymization is often applied to records prior to their analysis (*e.g.* date shifting, quantization, noise addition [50, 51]). Applying hospital-specific or probabilistic pseudonymization may break the equality of stratifying covariates among duplicates. Fourthly a subpopulation may be more subject to data duplication, thus inducing its over-representation in the validation fold. We moreover underline that the presented simulations are mostly illustrative: the size of biases measured in this article depend strongly on the data generation procedure, the model and the metric at stake.

Conclusion

When a model is trained and validated in a privacy-preserving federated learning setting, the presence of duplicated records may lead to over-optimistically biased estimates of its performances. We have shown that *stratified cross-validation* methodology can be used to avoid this over-optimistic bias without fully deduplicating records, although at the possible cost of a pessimistic bias that can be minimized by carefully choosing the stratifying covariate. We underline that *stratified cross-validation*, although of special importance in the case of federated learning where full deduplication is often unfeasible, also applies to the case of a centralized dataset with undetected duplicates and can therefore be used as an easy-to-implement sanity check. Although of possibly broad application, *stratified cross-validation* is only a partial solution to duplicate problems: full deduplication remains optimal to ensure databases integrity.

Acknowledgements

We thank the partners of "Healthchain" consortium for fruitfull discussions and E. Diard for her help with figure design.

Funding

This project is supported by Bpifrance as part of the "Healthchain" project, which resulted from the "Digital Investments Program for the major challenges of the future" RFP. As part of the "Healthchain" project, a consortium coordinated by Owkin (a private company) has been established, including the Substra association, Apricity (a private company), the Assistance Publique des Hôpitaux de Paris, the University Hospital Center of Nantes, the Léon Bérard Center, the French National Center for Scientific Research, the Ecole Polytechnique, the Institut Curie and the University of Paris.

Contributions

R.B. designed the methodology, conducted the simulations and wrote the manuscript.

R.G. and M.B. provided technical advice and manuscript feedbacks.

R.P. oversaw the project and helped with manuscript writing.

Conflict of interest statement

None.

Supplemental material

Code availability

The Jupyter notebook containing all the simulation code is available at https://doi.org/10.5281/zenodo.3614900 under Apache2.0 license.

Computation of the optimal accuracy

For each set of covariates $(x_1, x_2, ..., x_{10})$, the exact probability p_1 of a positive outcome $y = 1$ can be computed using equation (5). An ideal predictive model that optimizes its accuracy would predict $\tilde{y} = 1$ when $p_1 > 0.5$ and $\tilde{y} = 0$ otherwise. The average accuracy of such a model computed over the population of interest writes (see equations (2), (3), (4)):

$$Accuracy_{opt} = \int \Big[p_1 I(p_1 > 0.5) + (1 - p_1) I(p_1 < 0.5)\Big] \mathcal{N}\Big(\boldsymbol{\mu}, \Sigma\Big) dx_1 dx_2 ... dx_{10} \tag{6}$$

We estimate $Accuracy_{opt}$ through a Monte Carlo computation. We draw randomly 100000 records using the multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and we average the value of $p_1 I(p_1 > 0.5) + (1 - p_1) I(p_1 < 0.5)$ over the records. We measure an optimal accuracy of $Accuracy_{opt} = 0.88$.

Data visualization

In our simulations, records are generated following equations (2), (3), (4) and (5). Covariates are drawn from a multivariate Gaussian distribution (equation (2)) and outcomes are drawn randomly from the covariates using a logistic function (equation (5)). Figure 5 represents the distribution of positive outcomes relatively to covariates x_1, x_3, x_5, x_{10} obtained having drawn randomly 10000 records. These 4 covariates have been chosen arbitrarily in order to illustrate the structure of the datasets under scrutiny.

References

[1] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[2] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. Artificial intelligence in radiology. *Nat. Rev. Cancer*, 18(8):500–510, 2018.

[3] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.*, 24(11):1716–1720, 2018.

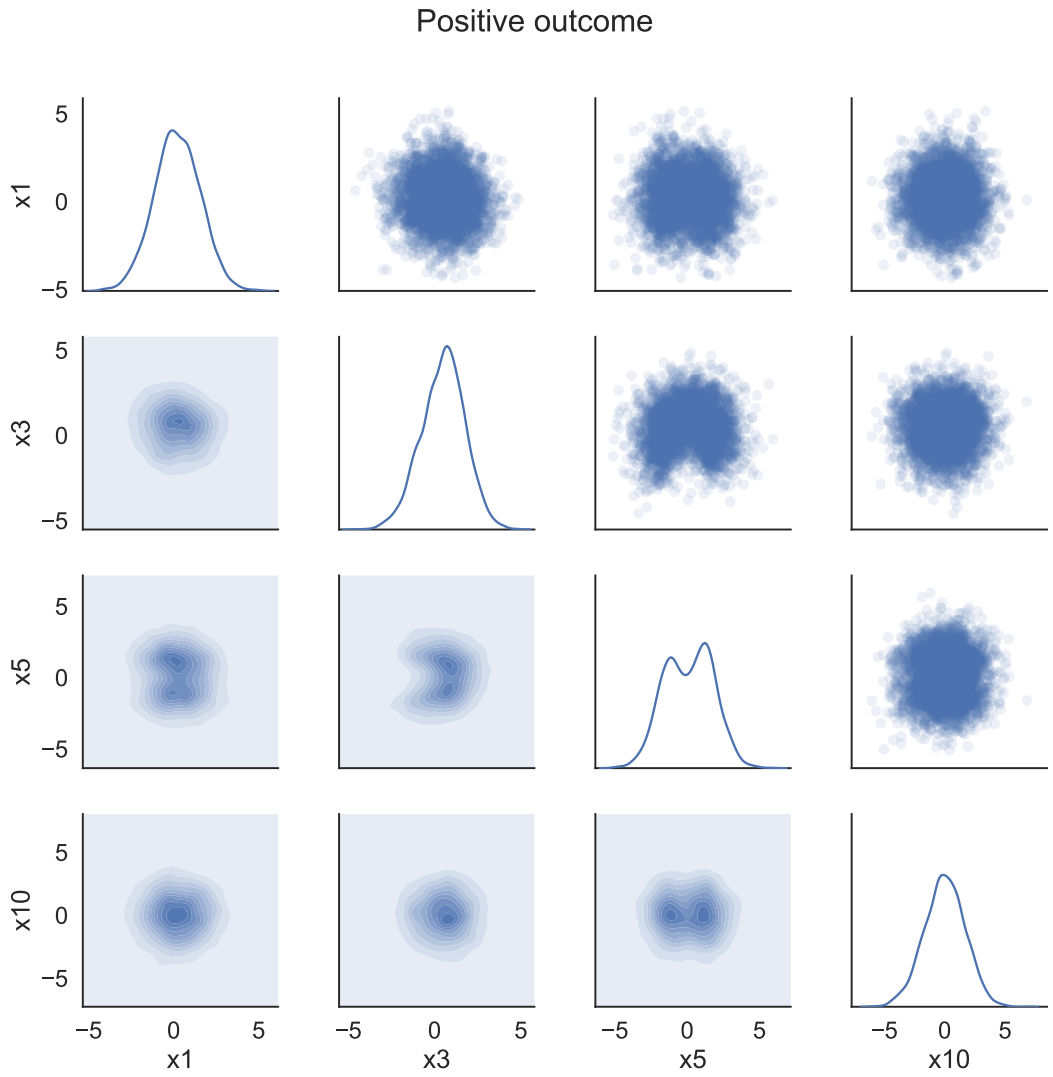

Figure 5: Diagonal: distributions of positive outcomes relatively to a single covariate. Upper-edge and lower-edge: scatter plots and density plots of positive outcomes relatively to two covariates.

[4] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*, 1:18, 2018.

[5] Fatemeh Rahimian, Gholamreza Salimi-Khorshidi, Amir H. Payberah, Jenny Tran, Roberto Ayala Solares, Francesca Raimondi, Milad Nazarzadeh, Dexter Canoy, and Kazem Rahimi. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med.*, 15(11):e1002695, 2018.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 2009.

[7] Tjeerd van der Ploeg, Peter C. Austin, and Ewout W. Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*, 14:137, December 2014.

[8] Julia Powles and Hal Hodson. Google DeepMind and healthcare in an age of algorithms. *Health Technol (Berl)*, 7(4):351–367, 2017.

[9] Fiona Caldicott. Review of data security, consent and opt-outs, 2016.

[10] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.*, 4(8):e1000167, August 2008.

[11] John Bohannon. Genetics. Genealogy databases enable naming of anonymous DNA donors. *Science*, 339(6117):262, January 2013.

[12] Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, January 2013.

[13] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun*, 10(1):3069, 2019.

[14] W. Nicholson Price and I. Glenn Cohen. Privacy in the age of medical big data. *Nat. Med.*, 25(1):37–43, 2019.

[15] Charu C. Aggarwal. On K-anonymity and the Curse of Dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 901–909. VLDB Endowment, 2005. event-place: Trondheim, Norway.

[16] Justin Brickell and Vitaly Shmatikov. The Cost of Privacy: Destruction of Data-mining Utility in Anonymized Data Publishing. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 70–78, New York, NY, USA, 2008. ACM. event-place: Las Vegas, Nevada, USA.

[17] Yves-Alexandre de Montjoye, Sébastien Gambs, Vincent Blondel, Geoffrey Canright, Nicolas de Cordes, Sébastien Deletaille, Kenth Engø-Monsen, Manuel Garcia-Herranz, Jake Kendall, Cameron Kerry, Gautier Krings, Emmanuel Letouzé, Miguel Luengo-Oroz, Nuria Oliver, Luc Rocher, Alex Rutherford, Zbigniew Smoreda, Jessica Steele, Erik Wetter, Alex Sandy Pentland, and Linus Bengtsson. On the privacy-conscientious use of mobile phone data. *Sci Data*, 5:180286, 2018.

[18] Joshua R. Vest and Kosali Simon. Hospitals' adoption of intra-system information exchange is negatively associated with inter-system information exchange. *J Am Med Inform Assoc*, 25(9):1189–1196, 2018.

[19] Yuan Wu, Xiaoqian Jiang, Jihoon Kim, and Lucila Ohno-Machado. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc*, 19(5):758–764, October 2012.

[20] Chia-Lun Lu, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, Xiaoqian Jiang, and Lucila Ohno-Machado. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc*, 22(6):1212–1219, November 2015.

[21] Reza Shokri and Vitaly Shmatikov. Privacy-Preserving Deep Learning. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1310–1321, New York, NY, USA, 2015. ACM. event-place: Denver, Colorado, USA.

[22] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *arXiv:1602.05629 [cs]*, February 2017. arXiv: 1602.05629.

[23] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 1175–1191, New York, NY, USA, 2017. ACM. event-place: Dallas, Texas, USA.

[24] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, December 2019. arXiv: 1912.04977.

[25] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards Federated Learning at Scale: System Design. *arXiv:1902.01046 [cs, stat]*, March 2019. arXiv: 1902.01046.

[26] Jean Louis Raisaro, Florian Tramèr, Zhanglong Ji, Diyue Bu, Yongan Zhao, Knox Carey, David Lloyd, Heidi Sofia, Dixie Baker, Paul Flicek, Suyash Shringarpure, Carlos Bustamante, Shuang Wang, Xiaoqian Jiang, Lucila Ohno-Machado, Haixu Tang, XiaoFeng Wang, and Jean-Pierre Hubaux. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *J Am Med Inform Assoc*, 24(4):799–805, July 2017.

[27] Jean Louis Raisaro, Juan Ramon Troncoso-Pastoriza, Mickael Misbach, Joao Sa Sousa, Sylvain Pradervand, Edoardo Missiaglia, Olivier Michielin, Bryan Ford, and Jean-Pierre Hubaux. MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. *IEEE/ACM Trans Comput Biol Bioinform*, 16(4):1328–1341, August 2019.

[28] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. *arXiv:1811.04017 [cs, stat]*, November 2018. arXiv: 1811.04017.

[29] Mathieu N. Galtier and Camille Marini. Substra: a framework for privacy-preserving, traceable and collaborative Machine Learning. *arXiv:1910.11567 [cs]*, October 2019. arXiv: 1910.11567.

[30] Rui Duan, Mary Regina Boland, Zixuan Liu, Yue Liu, Howard H. Chang, Hua Xu, Haitao Chu, Christopher H. Schmid, Christopher B. Forrest, John H. Holmes, Martijn J. Schuemie, Jesse A. Berlin, Jason H. Moore, and Yong Chen. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *J Am Med Inform Assoc*, December 2019.

[31] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. Big data. The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–1205, March 2014.

[32] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Sci Adv*, 4(1):eaao5580, 2018.

[33] Franz J. Király, Bilal Mateen, and Raphael Sonabend. NIPS - Not Even Wrong? A Systematic Review of Empirically Complete Demonstrations of Algorithmic Effectiveness in the Machine Learning and Artificial Intelligence Literature. *arXiv:1812.07519 [cs, stat]*, December 2018. arXiv: 1812.07519.

[34] Seong Ho Park and Kyunghwa Han. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*, 286(3):800–809, 2018.

[35] Sebastian Vollmer, Bilal A. Mateen, Gergo Bohner, Franz J. Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine S. L. McAllister, Puja Myles, David Granger, Mark Birse, Richard Branson, Karel GM Moons, Gary S. Collins, John P. A. Ioannidis, Chris Holmes, and Harry Hemingway. Machine learning and AI research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics and Effectiveness. *arXiv:1812.10404 [cs, stat]*, December 2018. arXiv: 1812.10404.

[36] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Trans. Knowl. Discov. Data*, 6(4):15:1–15:21, December 2012.

[37] Katie Harron, Angie Wade, Ruth Gilbert, Berit Muller-Pebody, and Harvey Goldstein. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med Res Methodol*, 14:36, March 2014.

[38] Wei Luo, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho, Svetha Venkatesh, and Michael Berk. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J. Med. Internet Res.*, 18(12):e323, 2016.

[39] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. The need to approximate the use-case in clinical machine learning. *Gigascience*, 6(5):1–9, 2017.

[40] Allison B. McCoy, Adam Wright, Michael G. Kahn, Jason S. Shapiro, Elmer Victor Bernstam, and Dean F. Sittig. Matching identifiers in electronic health records: implications for duplicate records and patient safety. *BMJ Qual Saf*, 22(3):219–224, March 2013.

[41] Jordan Everson and Julia Adler-Milstein. Gaps in health information exchange between hospitals that treat many shared patients. *J Am Med Inform Assoc*, 25(9):1114–1121, 2018.

[42] Katie Harron, Harvey Goldstein, and Chris Dibben. *Methodological Developments in Data Linkage*. John Wiley & Sons Inc., United States, 2015.

[43] Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Inform Syst*, 38(6):946–969, September 2013.

[44] Kassaye Yitbarek Yigzaw, Antonis Michalas, and Johan Gustav Bellika. Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation. *BMC Med Inform Decis Mak*, 17(1):1, 2017.

[45] Peeter Laud and Alisa Pankova. Privacy-preserving record linkage in large databases using secure multiparty computation. *BMC Med Genomics*, 11(Suppl 4):84, October 2018.

[46] N. A. Diamantidis, D. Karlis, and E. A. Giakoumakis. Unsupervised stratification of cross-validation for accuracy estimation. *Art. Int.*, 116(1):1–16, January 2000.

[47] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 785–794, 2016. arXiv: 1603.02754.

[48] Yang Liu, Zhuo Ma, Ximeng Liu, Siqi Ma, Surya Nepal, and Robert Deng. Boosting Privately: Privacy-Preserving Federated Extreme Boosting for Mobile Crowdsensing. *arXiv:1907.10218 [cs]*, July 2019. arXiv: 1907.10218.

[49] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, and Qiang Yang. SecureBoost: A Lossless Federated Learning Framework. *arXiv:1901.08755 [cs, stat]*, January 2019. arXiv: 1901.08755.

[50] Khaled El Emam and Luk Arbuckle. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly Media, Inc., December 2013. Google-Books-ID: 3RtRAgAAQBAJ.

[51] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.