A Decentralized Framework for Biostatistics and Privacy Concerns

Paul Mangold, Alexandre Filiot, Mouhamed Moussa, Vincent Sobanski, Grégoire Ficheur, Paul Andrey, Antoine Lamer

Integrated Citizen Centered Digital Health and Social Care
A. Värri et al. (Eds.)

doi:10.3233/SHTI200710

A Decentralized Framework for Biostatistics and Privacy Concerns

Paul MANGOLD[a,b,c,1], Alexandre FILIOT[a], Mouhamed MOUSSA[d], Vincent SOBANSKI[a], Gregoire FICHEUR[a,e], Paul ANDREY[a] and Antoine LAMER[e]

[a] *CHU Lille, INCLUDE: Integration Center of the Lille University Hospital for Data Exploration, 59000, Lille, France.*
[b] *INRIA Lille Nord Europe, Magnet Team, 59650, Villeneuve d'Ascq, France.*
[c] *ENS de Lyon, 69007, Lyon, France.*
[d] *CHU Lille, Pole d'Anesthesie-Reanimation, 59000, Lille, France.*
[e] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Evaluation des Technologies de sante et des Pratiques medicales, 59000, Lille, France.*

Abstract. Biostatistics and machine learning have been the cornerstone of a variety of recent developments in medicine. In order to gather large enough datasets, it is often necessary to set up multi-centric studies; yet, centralization of measurements can be difficult, either for practical, legal or ethical reasons. As an alternative, federated learning enables leveraging multiple centers' data without actually collating them. While existing works generally require a center to act as a leader and coordinate computations, we propose a fully decentralized framework where each center plays the same role. In this paper, we apply this framework to logistic regression, including confidence intervals computation. We test our algorithm on two distinct clinical datasets split among different centers, and show that it matches results from the centralized framework. In addition, we discuss possible privacy leaks and potential protection mechanisms, paving the way towards further research.

Keywords. federated learning, data privacy, biostatistics

1. Introduction

The advent of machine learning methods and the ongoing movement towards wide and high-quality data collection have made biostatistics a crucial component in medical research. Constituting large and representative datasets, which are mandatory either to have enough statistical power or to improve models' generalization, is not always feasible within a single medical center. A popular approach is thus to centralize data from multiple centers in one leading site and conduct the study there. With medical data, this centralization is often a practical challenge, as data is sensitive and must be handled within a controlled environment abiding by strong legal and ethical constraints.

An alternative approach, known as federated learning, consists in training statistical models in a decentralized way, leaving the data on each site, running computations

[1] Corresponding Author, Paul MANGOLD, INRIA Lille Nord Europe, 40 avenue Halley, 59650 Villeneuve d'Ascq, France; E-mail: paul.mangold@inria.fr

locally and communicating aggregated information between centers during the training phase. Such an approach has already been applied to medicine in a few studies, with the goal of preserving the privacy of sensitive data [1,2], as well as data owners' sovereignty.

This work takes a first step towards defining and implementing a decentralized learning framework for medicine, which differs from previous works in that it allows full decentralization, meaning that it does not require any center to play a central role in the computation (although the latter case remains an option). We aim at proving that this framework can produce results virtually identical to the ones obtained in a centralized setting on actual clinical data. To do so, we use two distinct datasets, fit logistic regressions and compute confidence intervals of the estimates. Finally, we put our work into perspective by highlighting some privacy concerns, together with privacy-preserving mechanisms that could address them, depending on desired privacy levels.

2. Methods

2.1. Decentralized Protocol for Logistic Regression

Logistic regression is fit by estimating the parameters that maximize the likelihood over the observed dataset. Iterative algorithms, such as gradient descent, are commonly used to do so. A decentralized version of gradient descent, as described in [3], can thus be used. In this setting, *each center* runs the following protocol:

- *Initialization.* Initialize local variables, and divide features by agreed-upon maximum values. This ensures faster convergence, without sharing private data. • *Training.* Iterate until convergence:
    - ∗ *Local Update.* Compute a local gradient and update local parameters.
    - ∗ *Communication and Aggregation.* Send local parameters to other centers and await theirs. Average the local and received parameters. Assign results as the new local parameters.
- *Confidence intervals computation.* Compute Fisher information on local dataset, and send it to others. Use these values to compute global confidence intervals.

Note that categorical variables are encoded as dummy variables, whose proper encoding requires either a set of agreed-upon values, or extra communications to determine those.

2.2. Datasets and Learning Scenarios

The first clinical dataset used in our experiments consists in measurements collected during caesarean sections performed at the Lille University Hospital. We aim at predicting fetal acidosis at birth based on six explanatory variables, including blood pressure drops during the operation. To simulate a multi-centric environment, the 775 records were randomly assigned to four equally-sized (up to one sample) chunks.

So as to provide reproducible results, the UCI heart disease dataset [4], available at https://archive.ics.uci.edu/ml/datasets/heart+Disease, was also used.
We aim at predicting the presence of a heart disease based on twelve explanatory variables, which mainly encompass clinical measurements at rest and during a controlled

physical effort. This data was collected in four distinct medical centers, with variable sample sizes (respectively 303, 261, 130 and 46 records, for a total of 740).

Three different learning scenarios are studied. The "centralized" scenario, in which the entire collated dataset can be used by a single center. The "all alone" scenario, in which each center tries to perform the study using only its local dataset. The "decentralized" scenario, in which centers communicate together without directly exchanging data records, following the protocol detailed in section 2.1.

To compare those three scenarios, we observe the estimated coefficients and their confidence intervals, checking whether they match, and if so, how precisely.

3. Results

3.1. Implementation

Our decentralized framework is implemented as a R package, available under the MIT license at https://gitlab.com/include-project/federate. The developed package handles network communications, and provides a way to simulate decentralized algorithms locally for testing purposes. It currently implements logistic regression with basic gradient descent, but may easily be expanded to comprise new algorithms, as only logical parts need to be re-implemented.

Algorithms are implemented in R with C++ integration using Rcpp. C++ libraries Armadillo and Asio are used for linear algebra and networking, with their respective R bindings RcppArmadillo and AsioHeaders, available at CRAN.

3.2. Experimental Results

The three learning scenarios were run on both the caesarean section and heart disease datasets. In the decentralized scenario, the algorithm is run for a few thousands iterations, inducing as many communication rounds. For both datasets, resulting odds ratios and confidence intervals are the same as in the centralized scenario up to 10^{-5} precision on each coefficient. Better precision may be achieved at the cost of more communication rounds and tuning. As for the all alone scenario, it results in unsatisfactory estimations in each center, due to insufficient population size. Odds ratios obtained in these various settings are shown in Figure 1. Clinical results on the caesarean dataset match those reported in [5]. Scripts for the heart disease dataset are provided in the git repository.

3.3. Privacy Concerns

Although decentralized machine learning naturally favors privacy preservation, keeping the data on site does not fully prevent sensitive information leaks. For federated deep learning, [6] show that shared information may reveal parts of the training dataset.

Since logistic regression shares the same underlying optimization procedure as deep learning, it may be vulnerable to similar attacks. This raises major concerns as local datasets are often small (e.g. for studies on rare diseases), the whole purpose being to gather enough data records to achieve statistical significance. Furthermore, individual

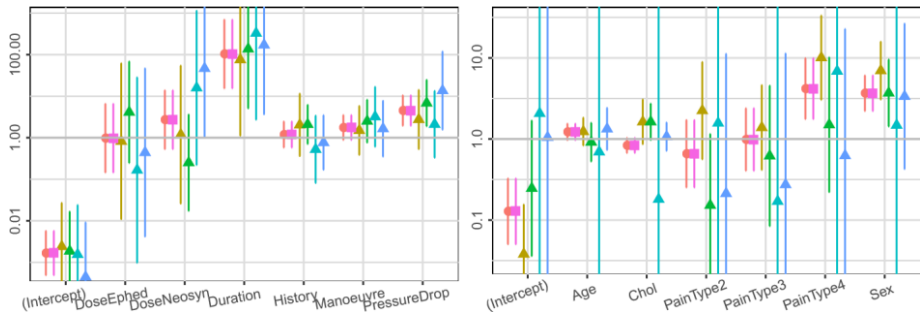

Figure 1. Odds ratios and 95% confidence intervals learned in the three distinct scenarios for the caesarean dataset (left, randomly uniformly split in 4), and heart dataset (right, split across the 4 actual sources, showing a subset of variables for readability). Circles, triangles and squares respectively represent the "centralized", "all alone" and "decentralized" scenarios. Each color represents odds ratio learned by a center. Horizontal grey line is 1, and confidence intervals not crossing it suggest a correlation exists between the variable and the outcome.

records are not the only sensitive information that may be revealed: local aggregated values, e.g. mortality rate, may be retrieved, which can expose centers' internal practices.

4. Discussions

4.1. Privacy Improvement Mechanisms

Local aggregates can be protected through secure aggregation [7], a protocol that consists in adding random masks on sent information that cancel out when computing the result. This yields an exact global average while preventing sent information from being revealed. It could be used during aggregation and confidence interval computations steps. However, this does not protect individual records from leaking. Differentially Private (DP) mechanisms [8] address this problem, by adding noise that blurs individual contribution on shared values, making it almost impossible to guess the presence of an individual in the dataset. This, however, widely impacts results' precision, and obtaining good accuracy while guaranteeing privacy generally requires very fine tuning of algorithms. Such mechanisms could be used at every communication step of our protocol, either before sending values (Local DP) or after their aggregation (Global DP), depending on trusted parties. Table 1 summarizes the impact of these mechanisms and describes who can infer information, thus requiring others' trust.

Table 1. Privacy mechanisms and who can infer what about records and aggregated values from local datasets.

Mechanism	Who can infer	Data Records	Local Aggregates	Precision
None	All	Not Protected	Not Protected	Exact
Sec. Agg.	All	Only Origin Protected	Protected	Exact
Global DP	Aggregator(s)	Protected	Not Protected	Inexact
Local DP	No one	Protected	Not Protected	Inexact
Sec. Agg. + DP	No one	Protected	Protected	Inexact

4.2. Perspectives

Our framework gives accurate results, echoing the conclusions of [2], within an acceptable number of communications rounds. Its modularity allows further experiments with more advanced optimization algorithms. It can also be extended to different learning tasks, including training deep neural networks, e.g. to classify medical images or learn word embeddings from hospital records. Besides, keeping data on site does not guarantee privacy. The latter should thus receive more attention in the future, notably by implementing secure aggregation and differentially private mechanisms. Precisely quantifying required privacy levels is mandatory to make informed choices of protection mechanisms. Therefore, a comprehensive study of effective data leakage appears to be the next step towards this direction. Full decentralization could further improve privacy, by enabling network topologies in which pairs of centers are distanced based on their mutual trust level. It may also lead to developing broader studies, directly leveraging measures from connected devices at patients' homes, or allow learning personalized parameters adjusted to local specificities.

5. Conclusion

Our decentralized framework gives very promising results, near-exactly matching those of the centralized scenario when fitting logistic regressions on two distinct clinical datasets. Its design and open-source implementation allow for its re-use, improvement and extension to other learning tasks. We have also identified a set of privacy-preserving mechanisms whose informed use can ease collaborations between clinical data holders.

References

[1] Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. Int J Radiat Oncol Biol Phys. 2017 Oct 1;99(2):344-352.

[2] Bogowicz M, Jochems A, Deist TM, et al. Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer. Sci Rep. 2020 Mar 11;10(1):4542.

[3] Lian X, Zhang C, Zhang H, Hsieh C, Zhang W, Liu J. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17); 2017 Dec 04-09; Long Beach, CA, USA.; 2017. p. 5336–46.

[4] UCI Machine Learning Repository [Internet]. Irvine (CA): Dua D, Graff C. 1988 - [cited 2020 Jul 22]. Available from: http://archive.ics.uci.edu/ml.

[5] Ngan Kee WD, Lee A. Multivariate analysis of factors associated with umbilical arterial pH and standard base excess after Caesarean section under spinal anaesthesia. Anaesthesia. 2003 Feb;58(2):125-30. doi: 10.1046/j.1365-2044.2003.02888.x. PMID: 12562407.

[6] Geiping J, et al. Near optimal neural network estimator for spectral¨ x-ray photon counting data with pileup. arXiv:2003.14053v1 [Preprint]. 2020 [cited 2020 Jul 22]: [26 p.]. Available from: https://arxiv.org/abs/2003.14053.

[7] Bonawitz K, Ivanov V, et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17); 2017 Oct 30-Nov 3; Dallas, TX, USA. 2017. p. 1175–91.

[8] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci. 2014 Aug;9(3–4):211-407. Available from: http://dx.doi.org/10.1561/0400000042