BY XIAOQIAN JIANG, PhD, SHUANG WANG, PhD,
STEPHANIE FEUDJIO FEUPE, LUCILA OHNO-MACHADO, PhD

PRIVACY-PROTECTING ANALYSIS OF DISTRIBUTED BIG DATA

A practical solution for sharing patient data while maintaining privacy protections.

Large clinical data research networks (e.g., PCORnet, HMORnet, ESPnet) have been established to accelerate scientific discovery and improve health. However, a big barrier to making full use of clinical data is the public's concern that researchers' access to demographics, diagnostic codes, genome sequences, etc., can pose risks for individual privacy, with potential implications for employment, security, and life and disability insurance. The current practice of "de-identifying" records before sharing them has limitations, including the likelihood of re-identification [1]. A better approach is to protect the privacy of patients involved in the study by controlling access to patient-level data in a way that respects their preferences while also facilitating research. This can be accomplished using customized distributed protocols that perform specific data analyses while storing and exchanging aggregated patient data through a trusted authority (TA)—an entity that can be trusted not to snoop into the data of the various parties.

Many existing algorithms for privacy-preserving distributed data analysis provide feasible but impractical solutions because they either involve very heavy computation (e.g., homomorphic encryption—computing on encrypted data) or introduce noise (e.g., methods based on differential privacy [2]). Other algorithms may have reasonable performance in a two-party setting but do not scale well to multiple parties.

One pragmatic and efficient framework for constructing accurate multivariate predictive models without ever exchanging patient-level data is Secure Multiparty Computing (SMC) with a TA. For example, biomedical computing centers with private HIPAA compliant clouds, such as iDASH (Integrating Data for Analysis, Anonymization and Sharing) [4] can offer themselves as a TA with whom authorized researchers can collaborate on distributed data analysis.

In this framework, local parties compute intermediary partial results (e.g., sufficient statistics, kernel matrices, etc.) and leverage the TA to combine partial results and coordinate iterative computation. This combination of partial results may be as simple as calculating a global average using partial averages and counts received from the parties, or may require the decomposition of algorithms in a way that allows combination of partial functions by the TA—for example by developing a distributed version of the Newton-Raphson algorithm [3]. Using a hub-and-spoke structure, local parties need only exchange information (at an aggregated level) with the TA through secure channels (such as through Secure Sockets Layer (SSL) in HTTPS) and always keep their patient-level data private.

This strategy has proven effective for a large family of data analysis models (including various generalized linear models and survival models) using different sets of patient data distributed across different servers [3] as well as using different sets of variables from the same patients distributed across different servers [5]—for example when patient phenotypes are hosted at a medical center and their genomes are hosted at a sequencing facility.

In addition to being privacy-protecting, this framework is efficient because the computation (1) is essentially parallelized (similar to the well known Map-Reduce architecture); (2) is amenable to optimization strategies such as prioritizing memory consumption or communication overhead; (3) creates a central point for building models, posing queries, and monitoring activities; (4) limits the need for communication between parties; and, perhaps most importantly, (5) ensures the reproducibility of experiments. □

REFERENCES:

1. Vaidya J, Shafiq B, Jiang X, *et al.* Identifying inference attacks against healthcare data repositories. In: *AMIA Summits Transl Sci Proc*. San Francisco, CA: 2013. 1–5.
2. Dwork C. Differential privacy. *Int Colloq Autom Lang Program* 2006;4052:1–12.
3. Wu Y, Jiang X, Kim J, *et al.* Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc* 2012;2012:758–64. doi:10.1136/amiajnl-2012-000862
4. Ohno-Machado L, Bafna V, Boxwala A a, *et al.* iDASH. Integrating data for analysis, anonymization, and sharing. *J Am Med Informatics Assoc* 2012;19:196–201.
5. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L, VERTIcal Grid lOgistic regression (VERTIGO), *J Am Med Inform Assoc* 2015;doi: 10.1093/jamia/ocv146.

DETAILS

Lucila Ohno-Machado is a professor of medicine and chair of the department of biomedical bioinformatics, and Xiaoqian Jiang and Shuang Wang are assistant professors in the biomedical informatics department at the University of California, San Diego, School of Medicine. Stephanie Feupe is a graduate student working in Ohno-Machado's lab.