12-26-2014

Development of a Web Service for Analysis in a Distributed Network

Xiaoqian Jiang
UC San Diego, x1jiang@ucsd.edu

Yuan Wu
Duke University, yuan.wu@duke.edu

Keith Marsolo
Cincinnati Children's Hospital Medical Center, keith.marsolo@cchmc.org

Lucila Ohno-Machado
University of California, San Diego, lohnomachado@ucsd.edu

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Development of a Web Service for Analysis in a Distributed Network

Abstract

Objective: We describe functional specifications and practicalities in the software development process for a web service that allows the construction of the multivariate logistic regression model, Grid Logistic Regression (GLORE), by aggregating partial estimates from distributed sites, with no exchange of patient-level data.

Background: We recently developed and published a web service for model construction and data analysis in a distributed environment. This recent paper provided an overview of the system that is useful for users, but included very few details that are relevant for biomedical informatics developers or network security personnel who may be interested in implementing this or similar systems. We focus here on how the system was conceived and implemented.

Methods: We followed a two-stage development approach by first implementing the backbone system and incrementally improving the user experience through interactions with potential users during the development. Our system went through various stages such as concept proof, algorithm validation, user interface development, and system testing. We used the Zoho Project management system to track tasks and milestones. We leveraged Google Code and Apache Subversion to share code among team members, and developed an applet-servlet architecture to support the cross platform deployment.

Discussion: During the development process, we encountered challenges such as Information Technology (IT) infrastructure gaps and limited team experience in user-interface design. We figured out solutions as well as enabling factors to support the translation of an innovative privacy-preserving, distributed modeling technology into a working prototype.

Conclusion: Using GLORE (a distributed model that we developed earlier) as a pilot example, we demonstrated the feasibility of building and integrating distributed modeling technology into a usable framework that can support privacy-preserving, distributed data analysis among researchers at geographically dispersed institutes.

Acknowledgements

X. Jiang, Y. Wu, and L. Ohno-Machado were funded in part by the Electronic Data Methods (EDM) Forum grant, NIH grants (R00LM011392, R21U54HL108460, UL1TR000100), and AHRQ grant R01HS019913. We thank Wenchao Jiang, Karapet Shaginyan, Asher Garland, Shuang Wang, Pinghao Li, Chialun Lu, Meng Xue for their contributions to this project. We thank Michele Day for generating Figure 2.

Keywords

privacy, distributed, predictive modeling

Disciplines

Health Information Technology | Health Services Research

Development of a Web Service for Analysis in a Distributed Network

Xiaoqian Jiang;[i] Yuan Wu;[ii] Keith Marsolo;[iii] Lucila Ohno-Machado[i]

Abstract

Objective: We describe functional specifications and practicalities in the software development process for a web service that allows the construction of the multivariate logistic regression model, Grid Logistic Regression (GLORE), by aggregating partial estimates from distributed sites, with no exchange of patient-level data.

Background: We recently developed and published a web service for model construction and data analysis in a distributed environment. This recent paper provided an overview of the system that is useful for users, but included very few details that are relevant for biomedical informatics developers or network security personnel who may be interested in implementing this or similar systems. We focus here on how the system was conceived and implemented.

Methods: We followed a two-stage development approach by first implementing the backbone system and incrementally improving the user experience through interactions with potential users during the development. Our system went through various stages such as concept proof, algorithm validation, user interface development, and system testing. We used the Zoho Project management system to track tasks and milestones. We leveraged Google Code and Apache Subversion to share code among team members, and developed an applet-servlet architecture to support the cross platform deployment.

Discussion: During the development process, we encountered challenges such as Information Technology (IT) infrastructure gaps and limited team experience in user-interface design. We figured out solutions as well as enabling factors to support the translation of an innovative privacy-preserving, distributed modeling technology into a working prototype.

Conclusion: Using GLORE (a distributed model that we developed earlier) as a pilot example, we demonstrated the feasibility of building and integrating distributed modeling technology into a usable framework that can support privacy-preserving, distributed data analysis among researchers at geographically dispersed institutes.

Introduction

A sufficient number of observations is needed to ensure the necessary power of a statistical test in a variety of studies (e.g., randomized clinical trials).[1] Thanks to the wide adoption of electronic health records (EHRs) and recent developments in distributed data networks (DRNs),[2–5] it is technically possible to pull together data from geographically distributed institutions to enable meaningful studies efficiently. This is particularly important for rare disease research. For example, United States researchers use clinical and Next-Generation Sequencing (NGS) data to study the Kawasaki disease, but local samples at Rady Children's Hospital (affiliated with the University of California, San Diego—UCSD) at San Diego are limited. International collaborations (e.g., with Japan and Singapore) can effectively promote data collection. But naïve procedures that involve patient-level data transferring will increase the privacy risk and may violate the institutional policy of sharing. As an example, unless explicitly requested and institutionally approved, all sensitive patient data must remain on the U.S. Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) project servers, and only aggregate data without personal identifiers may be transferred from VINCI. It is nontrivial to enable cross-institutional collaboration to respect patient privacy and institutional policies.[6-8]

Sweeney showed that it is possible to recover personal identifiable information (PII) through linkage of quasi-identifiers (e.g., ZIP codes, gender, and date of birth) from a private data set (e.g., hospital discharge data) to data from public data sets (e.g., demographics from a voter registry).[9] Other attack models showed that PII can be inferred under more restrictive conditions by using unforeseen auxiliary information.[10–12] We recently reviewed privacy technologies to support data sharing for comparative effectiveness research.[13] In this review, we explained that most synthetic data generation methods aim to protect privacy by introducing "indistinguishability" via the following: (1) *sanitizing* mechanisms like generalization or suppression of values,[14–16] and (2) synthetic

[i]University of California, San Diego, [ii]Duke University, [iii]Cincinnati Children's Hospital Medical Center

answer and data generation mechanisms that perturb the data.[17–20] These methods involve modifications of the original data and can make the disclosed data less useful for biomedical studies.

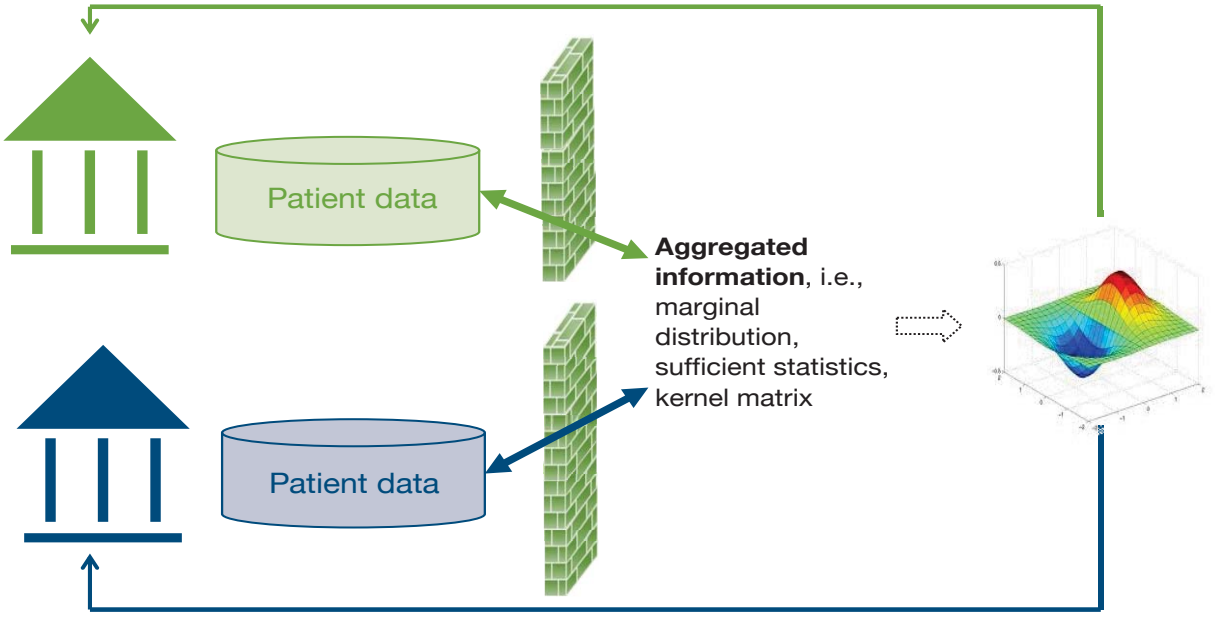
Another line of privacy technology research, which is not related to synthetic data releasing, aims at sharing models without sharing data through secure multiparty computing (SMC),[21–24] as illustrated in Figure 1. (SMC is a generalized method that enables secure data analysis across partitioned data from multiple institutes.) These approaches use security-enhanced protocols (e.g., transmission of partial statistics through encrypted protocols) to provide a practical solution for building accurate predictive models through cross-institutional collaboration. Several SMC-based privacy-preserving, data mining models such as linear regression,[25] clustering,[26] and support vector machine[27] have been developed to handle horizontally partitioned data (e.g., data of the same nature that are available at different institutions). Most relevant to our work are studies related to distributed logistic regression, which is the most popular predictive model in the biomedicine literature. Chu et al. introduced a map-reduce approach for logistic regression that enabled distributed model construction,[28] but it focused only on parallelization to accelerate the computation as this method arbitrarily allocates data and computation to a cluster of nodes to make the best use of available resources. Other articles[29,30] discussed privacy-preserving data analysis, but they did not explain how to evaluate the models, e.g., calculating the Hosmer-Lemeshow goodness-of-fit test (HL-test)[31] or areas under the receiver operating characteristic (ROC) curve,[32] in a distributed manner. These articles also did not provide a ready-to-use platform or software for biomedical researchers to easily deploy distributed learning frameworks in real health care environments.

In this article, we provide details on our recent work in developing collaborative software as a service (SAAS) for the Grid Logistic Regression (GLORE), previously published as "web service for Grid LOgistic Regression" (WebGLORE),[33] which is based on our previously established algorithm that showed how to build and evaluate a distributed logistic regression model.[34] We describe in detail the design and implementation of WebGLORE as an initial pilot effort to assess the feasibility of these methods in practice. We also discuss challenges and enabling factors in translating innovative privacy technology into a prototype system.

Methods

Imagine that an institution *A* wants to explore the contributing factors to the Kawasaki disease, which is a rare condition occurring in children that involves inflammation of the blood vessels. Because local observations are limited, *A* would like to collaborate with external partners (e.g., international research institutes) to build a logistic regression model using more cases to identify critical attributes more reliably. Due to the size of the data (each Next Generation Sequencing sample is about 200GB) and the privacy policy of participating institutes ("participants"), it is not feasible to move patient-level data around or build a centralized repository to meet this need. Our GLORE[35] framework, which allows global models to be constructed from distributed data sets without sharing raw data, can handle a use case like this. Our framework was motivated by the Scalable National Network for Effectiveness Research (SCANNER) project (http://scanner.ucsd.edu) and extended from a backbone program that runs on terminals to a web service, which does not require installation, to promote easy access and efficient collaboration in a distributed setting.

Figure 1. Privacy-Preserving Global Model Construction through Secure Multiparty Computing (SMC)



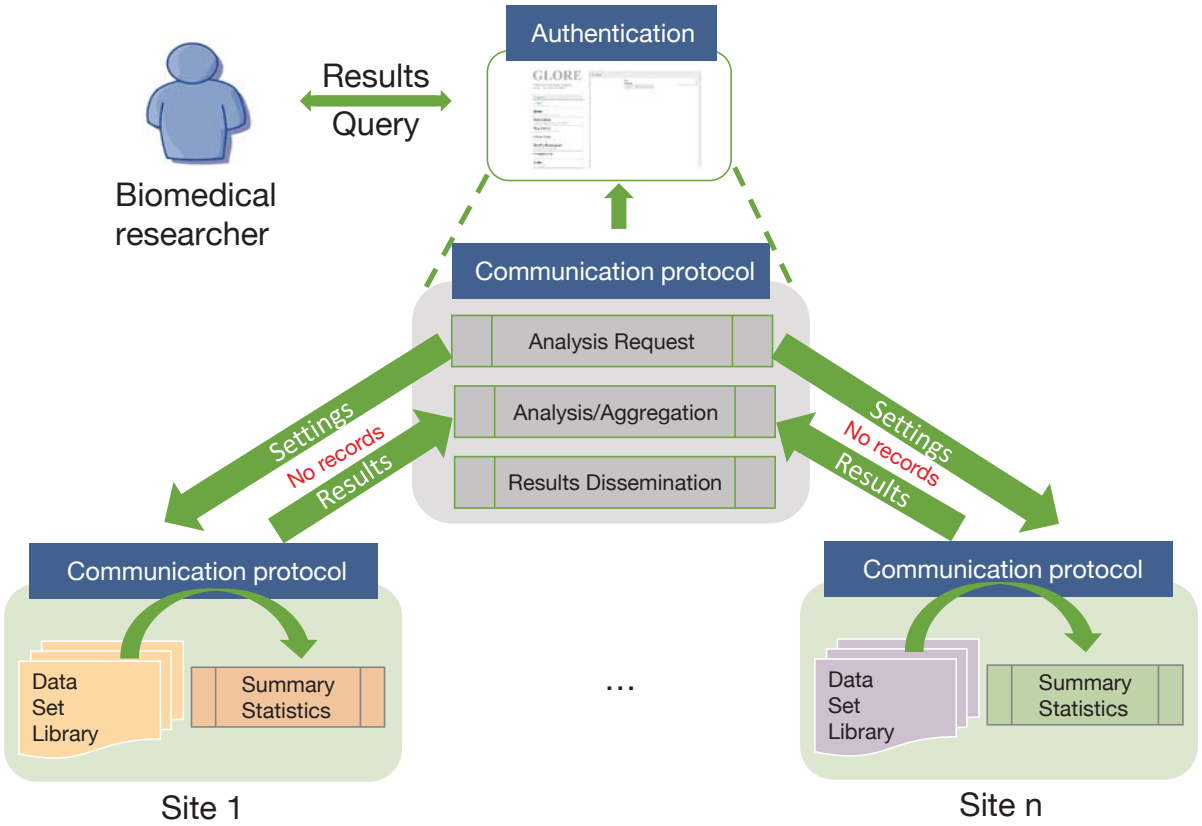
We designed the system to follow specifications elicited from multiple researchers at UCSD and researchers at Cincinnati Children's hospital. The web service was briefly described in an application note,[33] but it included very few details that are relevant for developers. We focus in this paper on how the system was conceived and implemented.

Figure 2 shows a high-level overview of the WebGLORE system. During the design and implementation of WebGLORE, we determined that the following requirements were important:

1. User friendliness: the system needed to be simple to use for biomedical researchers, with no requirements for installation or system configuration.
2. Platform independence: the system needed to be deployable in a variety of hosting environments (i.e., accommodate different operatin gsystems).
3. Security: we needed to ensure information security during information exchange.
4. Efficient evaluation: online summarization of global and local models needed to support quick comparisons.
5. Enforced authentication: only signed modules (Java applets and servlets) should interact with one another.

We followed a three-stage development approach by first implementing the backbone systems – R (https://www.dropbox.com/s/gmnrqgifdq9tjd7/glore_R.zip)and JAVA (https://code.google.com/p/glore/) and incrementally improving the user experience by continually interacting with potential users in different stages of the development. During the deployment process, we encountered challenges due to Information Technology (IT) infrastructure gaps such as platform compatibility, permission of third-party installation, JSP-Applet communication, etc. Our original design was a client-side graphical interface for the backbone system that communicated on specified ports, which turned out to be impractical due to a number of disadvantages: (1) some partners were reluctant to install third party software, (2) communication through prespecified ports brings security concerns, and (3) the application may not support all platforms unless implemented individually. These limitations were identified through discussions at an early stage of the development, which avoided extra effort to rectify them later. We ended up developing an applet-servlet architecture, which does not require installation or port specification, to support cross platform deployment and ease of use. We also developed the user interface to allow enough flexibility to meet the demands of potential users. Among various enabling factors for translating the innovative privacy-preserving, distributed modeling technology into a working prototype, we

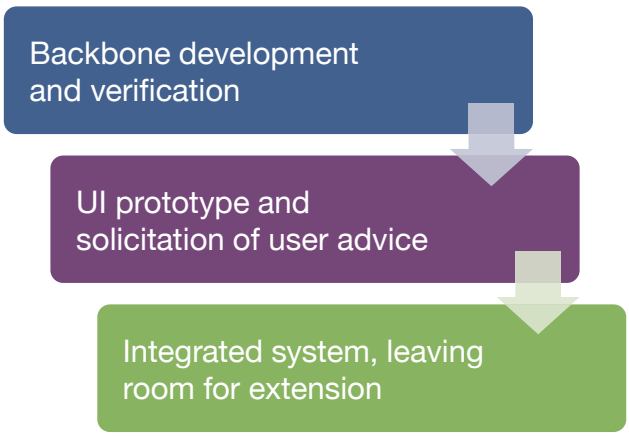
Figure 2. An Overview of the WebGLORE System



Note: The system leverages a secure communication protocol to interact with participating institutions to construct the global logistic regression model from distributed data sets. No records are exchanged or transmitted into a central node. The GLORE code communicated with participating institutes to get estimates from local data and aggregates them to build and evaluate a global model. [*Source:* Reproduced with permission from Kim et al., *JAMIA*[34]]

identified separating the interface development from algorithm development and integrated resource management (e.g., code, document, and calendar) to be most important. Specifically, we used the "Zoho Project management" (http://www.zoho.com/projects) to track tasks and milestones, and leveraged Google Code (https://code.google.com) and Apache Subversion (http://subversion.apache.org) (a free version control system) to share code among team members in the development phase.

Figure 3. Three-Stage Development Approach Used in the Implementation of the WebGLORE System



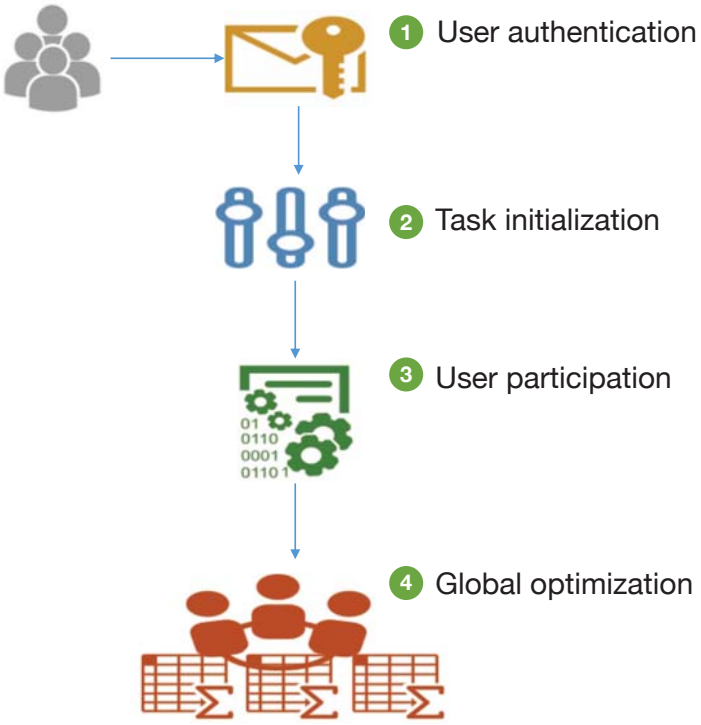
WebGLORE Function Specification

The pipeline of our collaborative framework for WebGLORE is illustrated in Figure 4. We developed four major modules to handle *user registration*, *initiator task creation*, *user participation*, and *collaborative model construction*, respectively. Dynamic content updates and synchronized model optimization are at the core of our system. The rest of this section will elaborate the functionalities that are involved.

Figure 4 shows the pipeline of WebGLORE at a high level, which includes four major components (user authentication, task initialization, user participation, and global optimization). These components are elaborated in details Figure 5, which illustrates the workflow from a user's perspective in detail.

WebGLORE first checks if a user is registered or not using its user authentication component (❷). The next component to be activated after authentication is the task initiation component (❸), which starts with specifying the parameters (name, expiration date, participants' email address, as well as the path of local training data to be linked). Note that the local computation module (an Applet resides in the Java Virtual Machine) links to the local data will compute necessary intermediary statistics to be exchanged but no patient-level data leave the host. An important challenge in *task initialization* is to ensure the consistency of predictive attributes across participants because a single inconsistent attribute can jeopardize the entire learning process. Ideally, data should match syntactically (i.e., name consistency) and semantically (i.e., value consistency). Our system supports syntactic validation by checking whether attribute names (specified by the

Figure 4. High-Level Pipeline of the Collaborative Framework of WebGLORE



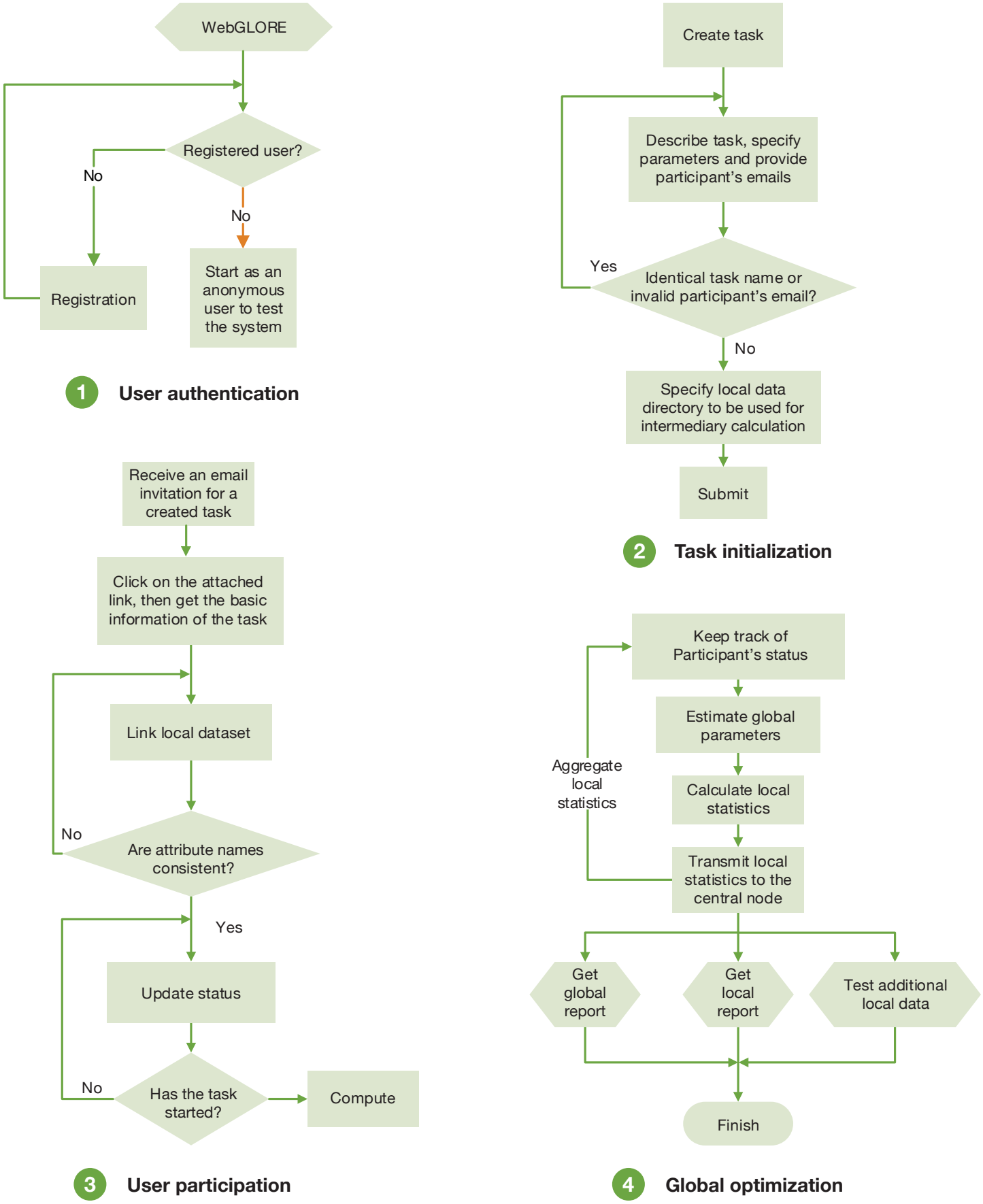
first row of local data files in our convention) are consistent to support basic data harmonization. We do not support semantic validation at this point. After passing the syntactic consistency check, the user-initiator can invite participants, who have the same type of data (i.e., same attributes) and want to build a global logistic regression.

After a task is initialized, the user participation component (❹) will be activated. Each invited participant will receive an email from the GLORE server; each one is provided with a unique link (e.g., hashed by task name and participant's email). The system also memorizes existing collaborators and stores their emails (in a list for quick retrieval from the invitation panel). After participants confirm attendance and the task initiator triggers "start calculating", the server begins to interact with all participants, including the task initiator. The system is also responsible for tracking user status (i.e., online/offline), model learning progress, as well as the value of the log-likelihood (i.e., the objective function) at each step during the computation.

When sufficient participants accepted the invitation and got their systems ready (i.e., completed the data linkage process), the task-initiator can trigger the global computation component (❺). Participants who miss the collaborative model construction will receive a notification. When the computation starts, it is synchronized across all participants.

WebGLORE will first assign random values to global parameters (for the predictors). Then, local statistics (i.e., gradient, Hessian matrix) will be calculated using these global parameters based on

Figure 5. Detailed Workflow of WebGLORE That Illustrates Its Major Components



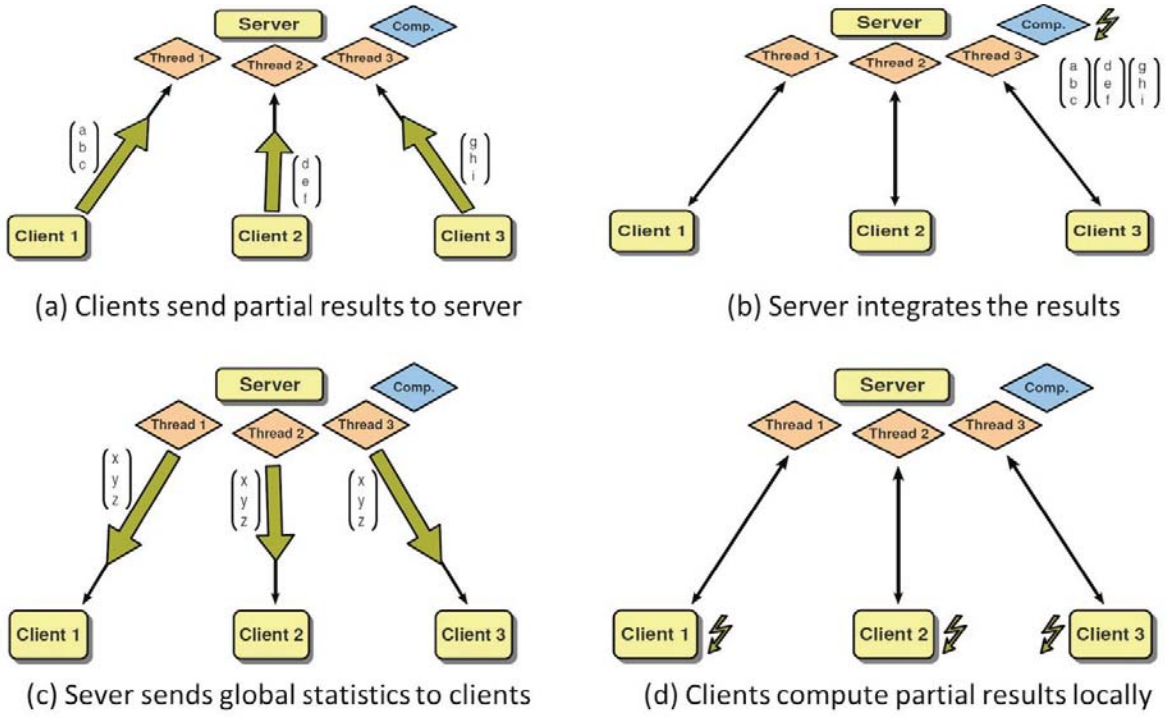
local datasets. The local statistics are transmitted to the central node, which combines the gradients and the Hessian matrix to conduct one iteration of the Newton Ralphson algorithm for optimization. This process is iterated in the distributed manner until global parameters converge or the maximum number of iterations is reached (http://jamia.bmj.com/content/suppl/2012/04/16/amiajnl-2012-000862.DC1/Appendix.pdf). The computation process is transparent to participants so they can watch inputs and outputs at each step (tracked in a panel on the computing web page).

When the calculation is finished, each participant will receive separate reports that allow he or she to compare the global model to locally trained logistic regression models. Note that the global report will only be created once (automatically by the WebGLORE system) for each task, and it can be retrieved through a unique URL that is accessible only by the task participants. For local reports, individual users click the button "Get local report" that triggers a one-time calculation using the linked local data. There is no information exchange in this step and the computation takes place in the browser (through the Java Virtual Machine applet). Each report contains important statistics, graphs, and performance measures such as sensitivity, specificity, Area under the ROC curve (AUC), and Hosmer-Lemeshow test p-values. The report also shows the plots of the ROC curve and of the reliability diagram. WebGLORE is open source, and the code is freely available at https://code.google.com/p/webglore/.

On the WebGLORE website, we embedded three short demo videos (http://dbmi-engine.ucsd.edu/webglore2/instructions.html) to help first time users. The first one is an introduction of WebGLORE, which walks users through different pages and the authentication process. The second demo is to show how WebGLORE can be used as a standard online logistic regression package, which works for a single user case. The last demo illustrates using WebGLORE to initiate and execute cross-institution collaborations.

The development of WebGLORE is an experience of constantly changing design and architecture to meet the need of potential partners. Our ImprovedCareNow! collaborator provided valuable suggestions to build a secure and easy-to-deploy framework, amenable to the existing healthcare environment. Our research partner from Duke University commented that comprehensive statistical reports for both local and global models should be provided to facilitate users in conducting comparative effectiveness research. Yet another helpful instruction was from EDM forum to emphasize on user experience in research computing. All these suggestions are highly valuable and we took them into serious consideration to redesign the system in multiple iterations. As a result, our package evolved from a port-based command-line interface to a heavy-client light-server webservice architecture, which includes necessary elements to meet the suggestions. In specific, WebGLORE carry out computation tasks locally to ensure data privacy and leverages the server to combine results for global analysis. We developed signed communication protocols between servlets and applets using Java, as illustrated in Figure 5. The applets are embedded in web pages (and executed by the Java Virtual machine) to handle local computation so that data never leave their host institutions. Since only signed applets can execute and communicate with servlets, we can easily check the validity of inputs from participants on the server side.

Figure 6. The Servlet and Applet Architecture Used in the GLORE System



Notes: The server generates one thread for each client request and synthesizes intermediary estimates used in the distributed Newton-Raphson algorithm. Note that (a,b,c) stands for intermediary results associated with gradients and variance matrices while (β_1, ... , $\beta\kappa$) stands for parameters.

We validated the system using public- and private data sets, which shows the same results as the centralized model. For more details, please refer to the Appendix for our system validation.

Discussion and Conclusion

We were able to construct a web service for global logistic regression modeling using distributed data sets that followed design specifications elicited from multiple researchers at UCSD and a collaborator at another institution. Because the information exchange required to build the model does not involve transmission of patient-level data, our system has reduced privacy risk when compared to the more common centralization of data followed by local computation of parameters. In addition, the WebGLORE reduces computation time thanks to the parallelization with multiple clients.

There are a number of statistical and practical limitations in WebGLORE. For example, it does not handle missing values or attribute domain mismatch (i.e., when the same attribute has different values in different data sets). Another limitation is that our system does not handle local site effects (fixed or random). Because these effects have the potential to significantly influence results of the analyses, we plan to address them in future work. For a handful of participants, it may be viable to assume fixed cluster effects and learn parameters for additional binary attributes that can represent the unknown factors associated with individual institutions. Random effects are more difficult to account for, especially in a distributed setting, and thus will require further investigation. While WebGLORE has limitations, it shows great promise in obtaining accurate analysis results in a distributed and privacy-preserving manner through exchanging aggregated partial estimates across the Information-Centric Networking (ICN).

The current version of WebGLORE still has some practical limitations. It interacts with local data and conducts local computation through the JAVA applet application, which requires certificates from a trusted authority to ensure security (starting with Java 7 update 51 in 2014, self-signed certificate is blocked by default). In addition, some users showed concerns about allowing automated scripts to access sensitive data and transfer information even though the process is transparent. To tackle these concerns, we plan to improve our implementation by supporting asynchronized human-in-the-loop authentication during every information-transmission step, and we plan to migrate the next version of WebGLORE to Chrome Apps, which is capable as a native application, but as safe as a web page, to mitigate the security issue.

Innovation in biomedicine requires health care researchers to constantly incorporate and customize state-of-the-art technologies into the available IT infrastructure. To make these translational efforts meaningful with a minimal burden to the end users (researchers, clinicians, patients), it is important to expose novel functionality in a natural and intuitive manner. Using GLORE as the basis, we showed how we built and integrated various technologies into a usable framework to support a distributed privacy-preserving modeling. We would like to find more biomedical applications for WebGLORE, and we welcome collaborations in extending it to general purpose, distributed data-analysis applications through the development of a wide range of secure primitives for partition data.

Acknowledgments

X. Jiang, Y. Wu, and L. Ohno-Machado were funded in part by the Electronic Data Methods (EDM) Forum grant, NIH grants (R00LM011392, R21U54HL108460, UL1TR000100), and AHRQ grant R01HS019913. We thank Wenchao Jiang, Karapet Shaginyan, Asher Garland, Shuang Wang, Pinghao Li, Chialun Lu, Meng Xue for their contributions to this project. We thank Michele Day for generating Figure 2.

Contributors

Xiaoqian Jiang led the development of the web service, wrote major portions of the manuscript, conducted experiments, and produced most figures. Yuan Wu co-led the development of the web service and contributed to the methodology. Keith Marsolo generated results associated with the ImprovedCareNow! network and provided useful design specifications. Lucila Ohno-Machado guided this study and provided thorough editing.

References

1. Moher D, Dulberg CS, Wells G a. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;**272**:122–4.
2. Schilling L, Kwan BM, Drolshagen CT, *et al*. calable Architecture for Federated Translational Inquiries Network (SAFTINet) tTechnology iInfrastructure for a dDistributed dData nNetwork. *eGEMs* 2013;**1**:1–13.
3. Holbrook AM, Keshavjee K, Troyan S. Designing a national eHealth dDrug sSafety and eEffectiveness rResearch nNetwork. *J Popul Ther Clin Pharmacol* 2011;**18 (2)**:e196–e197.
4. Brown JS, Holmes JH, Shah K, *et al*. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care* 2010;**48**:S45–S51.
5. Maro JC, Platt R, Holmes JH, *et al*. Design of a national distributed health data network. *Ann Intern Med* 2009;**151**:341–4.
6. Malin BA, Sweeney LA. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J Biomed Informatics* 2004;**37**:179–92.
7. McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. *J Am Med Inform Assoc* 2013;**20**:29–34.
8. Andrews EB. Data privacy, medical record confidentiality, and research in the interest of public health. *Pharmacoepidemiol Drug Saf* 1999;**8**:247–60.
9. Sweeney L. Uniqueness of simple demographics in the US population. *LIDAP-WP4 Carnegie Mellon Univ Lab Int Data Privacy, Pittsburgh, PA* 2000.

10. Ganta SR, Kasiviswanathan SP, Smith A. Composition attacks and auxiliary information in data privacy. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*. New York, New York, USA: : ACM Press 2008. 265–74.
11. Kifer D. Attacks on privacy and deFinetti's theorem. In: *Proceedings of the 35th SIGMOD international conference on Management of data*. New York, NY, USA: : ACM 2009. 127–38.
12. Wong RC-WW, Fu AW-CC, Wang K, *et al*. Can the utility of anonymized data be used for privacy breaches? *ACM Trans Knowl Discov Data* 2011;**5**:1–24.
13. Jiang X, Sarwate AD, Ohno-Machado L. Privacy technology to support data sharing for comparative effectiveness research: a systematic review. *Med Care* 2013;**51**:S58–65.
14. Trombetta A, Jiang W, Bertino E, *et al*. Privately updating suppression and generalization based k-anonymous databases. In: *Proc. of the 24th International Conference on Data Engineering (ICDE)*. Cancún, México: 2008. 1370–2.
15. Wang K, Yu PS, Chakraborty S. Bottom-up generalization: a data mining solution to privacy protection. In: *Proc. of the 4th IEEE International Conference on Data Mining (ICDM)*. Brighton, UK: 2004. 249 – 256.
16. Li T, Li N. Towards optimal k-anonymization. *Data Knowl Eng* 2008;**65**:22–39.
17. Dwork C. Differential privacy. *Int Colloq Autom Lang Program* 2006;**4052**:1–12.
18. Kargupta H, Datta S, Wang Q, *et al*. Random-data perturbation techniques and privacypreserving data mining. *Knowl Inf Syst* 2004;**7**:387–414.
19. Chen K, Liu L. Geometric data perturbation for privacy preserving outsourced data mining. *Knowl Inf Syst* 2011;**29**:657–95.
20. Chaytor R, Wang K, Brantingham P. Fine-grain perturbation for privacy preserving data publishing. In: *Proc. of the 9th IEEE International Conference on Data Mining (ICDM)*. Miami, FL: 2009. 740–5.
21. Teng Z, Du W. A hybrid multi-group approach for privacy-preserving data mining. *Knowl Inf Syst* 2009;**19**:133–57.
22. Huang Y, Lu Z, Hu H. Privacy preserving association rule mining with scalar product. In: *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*. 2005. 750–5.
23. Yu H, Jiang X, Vaidya J. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In: *Proceedings of the ACM symposium on Applied computing (SAC '06)*. New York, NY: 2006. 603–10.
24. Wu Y, Jiang X, Ohno-machado L. Preserving Institutional Privacy in Distributed Binary Logistic Regression. In: *AMIA Annu Symp*. Chicago, IL: 2012. 1450–8.
25. Hall R, Fienberg S, Nardi Y. Secure multiple linear regression based on homomorphic encryption. *J Off Stat* 2011;**27**:669–91.
26. Wei-jiang X, Liu-sheng H, Yong-long L, *et al*. Privacy-preserving DBSCAN clustering Over vertically partitioned data. In: *Multimedia and Ubiquitous Engineering, 2007. MUE '07. International Conference on*. 2007. 850–6.
27. Vaidya J, Yu H, Jiang X. Privacy-preserving SVM classification. *Knowl Inf Syst* 2008;**14**:161–78.
28. Chu CT, Kim SK, Lin YA, *et al*. Map-reduce for machine learning on multicore. *Adv Neural Inf Process Syst* 2007;**19**:281–8.
29. Slavkovic AB, Nardi Y, Tibbits MM. "Secure" Logistic regression of horizontally and vertically partitioned distributed databases. In: *Seventh IEEE International Conference on Data Mining Workshops (ICDMW)*. Omaha, NE: 2007. 723–8.
30. El Emam K, Samet S, Arbuckle L, *et al*. A secure distributed logistic regression protocol for the detection of rare adverse drug events. *J Am Med Inform Assoc* 2013;**20**:453–61.
31. Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: Wiley-Interscience 2000.
32. Lasko TA, Bhagwat JG, Zou KH, *et al*. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform* 2005;**38**:404–15.
33. Jiang W, Li P, Wang S, *et al*. WebGLORE: a Web service for Grid LOgistic REgression. *Bioinformatics* 2013;[Epub ahead of print].
34. Kim KK, Browe DK, Logan HC, *et al*. Data governance requirements for distributed clinical research networks: triangulating perspectives of diverse stakeholders. *J Am Med Inform Assoc* 2013;**21**:714–9.
35. Wu Y, Jiang X, Kim J, *et al*. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Informatics Assoc* 2012;**19**:758–64.
36. Wolberg WH, Street WN, Heisey DM, *et al*. Computer-derived nuclear features distinguish malignant from benign breast cytology. *Hum Pathol* 1995;**26**:792–6.
37. Frank A, Asuncion A. UCI Machine Learning Repository. 2010.http://archive.ics.uci.edu/ml
38. Crandall W, Kappelman MD, Colletti RB, *et al*. ImproveCareNow: The development of a pediatric inflammatory bowel disease improvement network. *Inflamm Bowel Dis* 2011;**17**:450–7.

Appendix A. System Validation

We used one public data set and one private clinical data set to validate the model. The public one is the Wisconsin breast cancer (diagnostic) data set[36], which was downloaded from the University of California Irvine (UCI) machine learning repository[37] from the public URL. When the rows with missing values are removed, this data set has 683 records, 9 features, and 1 target variable indicating a benign (class=0) or malignant (class=1) tumor. We split the data into 342 and 341 records to represent two sites.

The private data set is from the ImproveCareNow Network, which was collected for the purpose of improving the outcomes of care for children with inflammatory bowel disease (IBD).[38] Two sites with 20 patients and 47 patients, respectively, participated. Multiple observations were associated with individual patients in the original data. There were 9 predictor attributes and 1 target variable, indicating whether the patient responded to treatment. The data set used in this demonstration project contained only one observation per patient. We first conducted the experiment using the Wisconsin breast cancer data set. The model converged very quickly (fewer than 8 iterations to achieve a tolerance of 10^-6). The results of GLORE and LR (locally trained with combined data sets) are the same, as illustrated in Table A1. We also get the same goodness-of-fit (HL-test based on deciles p=0.06) and discrimination result (AUC=0.996).

WebGLORE also generates the ROC curve and reliability diagram to illustrate the discrimination and calibration of a trained model visually. Figure A1 shows an example.

Our next experiment was conducted on the ImproveCareNow Data. Table A2 lists the estimated parameters and their statistics.

Figure A2 shows the corresponding ROC curve and reliability diagram of the distributed model, for which the AUC is 0.753 and the p-value of the HL-test based on deciles is 0.13.

Table A1. GLORE Estimated Parameters and Statistics Using the Wisconsin Breast Cancer Data Set

Predictor	Estimate	Std. Error	Z-value	Pr(>|z|)
Intercept	-10.1039	1.1749	-8.5999	<0.0001
Clump Thickness	0.535	0.142	3.7672	0.0002
Uniformity of Cell Size	-0.0063	0.2091	-0.03	0.976
Uniformity of Cell Shape	0.3227	0.2306	1.3994	0.1617
Marginal Adhesion	0.3306	0.1235	2.6783	0.0074
Single Epithelial Cell Size	0.0966	0.1566	0.6171	0.5372
Bare Nuclei	0.383	0.0938	4.0815	<0.0001
Bland Chromatin	0.4472	0.1714	2.6093	0.0091
Normal Nucleoli	0.213	0.1129	1.8873	0.0591
Mitoses	0.5348	0.3288	1.6267	0.1038

Figure A1. ROC Plot and Reliability Diagram Generated Using the Wisconsin Breast Cancer Data Set

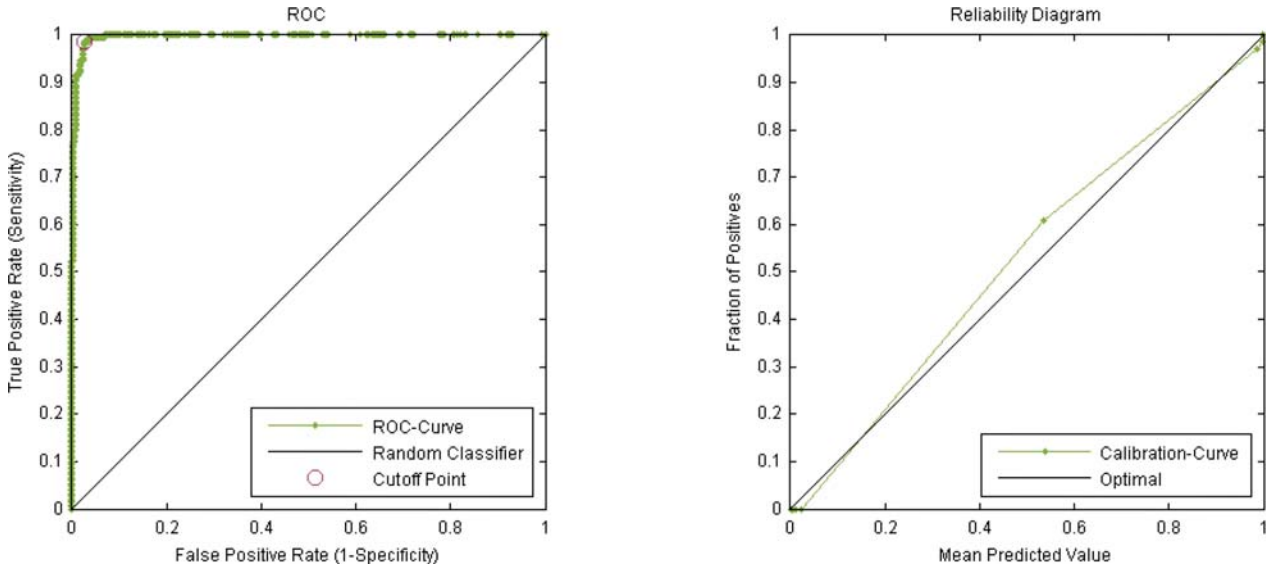


Table A2. GLORE Estimated Parameters and Statistics Using the ImproveCareNow Data Set

	Estimate	Std. Error	Z-value	Pr(>|z|)
Intercept	-1.369	3.3874	-0.4124	0.6801
Patient on biologics	0.7773	1.0627	0.7314	0.4645
Days since diagnosis	0.0002	0.0006	0.3075	0.7585
Gender	-0.4021	0.9262	-0.4342	0.6642
Race	0.2650	0.3983	0.665	0.5058
Age in years at start of treatment	-0.0234	0.1489	-0.1572	0.8751
Extent of disease	0.0893	0.1409	0.6336	0.5263
Patient on thiopurine	0.7574	0.6623	1.1437	0.2527
Patient on methotrexate	0.0000	3162.2777	0.0000	1.0000
Patient on salicylate	1.9536	1.1546	1.6919	0.0907
Patient on steroids	0.8684	0.6580	1.3197	0.1869

Figure A2. ROC Plot and Reliability Diagram Generated Using the ImproveCareNow Data Set