A distributed multi-site latent class analysis (dMLCA) algorithm for federated disease subphenotype detection

Naimin Jing[1,10], Xiaokang Liu[1], Qiong Wu[1], Suchitra Rao[2], Asuncion Mejias[3], Mitchell Maltenfort[4], Julia Schuchard[4], Vitaly Lorman[4], Hanieh Razzaghi[4], Ryan Webb[4], Chuan Zhou[5], Ravi Jhaveri[6], Grace M. Lee[7], Nathan M. Pajor[8], Deepika Thacker[9], L. Charles Bailey[4], Christopher B. Forrest[4], Yong Chen[1]

[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

[2]Department of Pediatrics, University of Colorado School of Medicine and Children's Hospital Colorado, Aurora, CO

[3]Division of Infectious Diseases, Department of Pediatrics, Nationwide Children's Hospital and The Ohio State University, Columbus, OH

[4]Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA

[5]Center for Child Health, Behavior and Development, Seattle Children's Hospital, Seattle, WA

[6]Division of Infectious Diseases, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL

[7]Department of Pediatrics (Infectious Diseases), Stanford University School of Medicine, Stanford, CA

[8]Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, OH

[9]Division of Cardiology, Nemours Children's Health, Wilmington, DE

[10]Biostatistics and Research Decision Sciences, Merck & Co., Inc, Kenilworth, NJ 07033, USA

Background

Large data networks emerge recently to facilitate collaborative learning across multiple institutions for increased modeling performance. However, data sharing prohibition and the heterogeneity in population among institutions pose new challenges in data modeling algorithms. Unlike traditional settings, the data are distributed across institutions instead of centralized and only summary-level statistics can be shared. There have been many efforts in developing federated learning algorithms for distributed data, see [1-16]. However, the above-mentioned methods all focused on supervised learning, while less attention paid to unsupervised clustering tasks needed for disease subphenotyping.

Multisystem inflammatory syndrome in children (MIS-C) is a form of serious post-acute sequelae of SARS-CoV-2 infection (PASC) in children. The clinical features of MIS-C are diverse and complex making the diagnosis of MIS-C difficult. Therefore, it is essential to characterize its disease patterns by subphenotypes for improved recognition and treatment. Due to the rareness of MIS-C, data integration across multiple hospitals is essential to a reliable result. Latent class analysis (LCA) is a statistical model to detecting disease subphenotypes but its implementation in a multi-site setting is challenging.

The first challenge is that the commonly used divide-and-conquer meta-analysis-type methods cannot be applied. Divide-and-conquer methods obtain an estimator from each site individually and then combine the estimators according to a certain rule (e.g., average). However, the LCA models trained from different sites cannot be combined because applying LCA separately at different sites may not yield the same set of latent classes due to the unsupervised nature of LCA. Such inconsistency can lead to ambiguity in matching latent classes among sites. Another challenge is the commonly observed existence of heterogeneity among patient populations from different sites due to varied geographic regions, community referral patterns, and health system structures and processes. Current works on multi-site LCA simply pooled all the data together and treated it as a single-site analysis without interpreting the heterogeneity across sites [17,18], which may lead to biased estimations [19, 20].

Motivated by the clinical needs for MIS-C subphenotying and the lack of federated subphenotyping methods, we aimed to develop a distributed multisite latent class analysis (dMLCA) that properly accounts for the between-site population heterogeneity and requires no individual-level data sharing across institutions while achieving the same results as using centralized data. We then demonstrated the usage of our method by applying it to a dataset of MIS-C patients from nine PEDSnet institutions.

Methods

We proposed a new model formulation for dMLCA based on the traditional LCA. To solve the first challenge of class-matching difficulty, we enforce the latent class characteristics to be the same across sites. This is reasonable because disease subtypes depend on the mechanism of the disease not sites. To deal with the population heterogeneity, we let the proportion of each subphenotypes to be different across sites. Then the distribution of a categorical manifest variable $y = (y_1, y_2, \dots, y_q)$ from site k can be expressed by

$$f_k(y) = \sum_{c=1}^{C} \lambda_{kc} f(y, \pi_c),$$

where λ_{kc} is the proportion of subphenotype c on site k and π_c is the mean of y. We obtain the estimation of parameters (θ) by maximizing likelihood through EM algorithm. At the t-th iteration of EM, the Newton-Raphson algorithm is applied to solve a target function $Q(\theta|\theta^{t-1})$.

A key observation to handling the data-sharing prohibition is that the updating formulas in EM algorithms are decomposable by sites. Therefore, at each iteration, each site only needs to calculate and communicate the decomposable part using its local data and transfer the results to the lead site to update the estimation. No patient-level data sharing is needed. For better communication efficiency, we set the number of iterations in the Newton-Raphson algorithm to 1. Since the updating formulas are exactly calculated, our method is lossless compared with LCA on centralized data.

We then applied the dLMCA algorithm to the data of MIS-C patients from 9 institutions between March 2020 and December 2021, including 864 children and adolescents < 21 years of age, to

detect subphenotypes of MIS-C. The data were harmonized by PEDSnet Common Data Model, which was developed based on OMOP CDM. The data in this study was from multiple sites but accessible from one site.

Results

dMLCA separated the complex MIS-C cohort into three clinically interpretable subphenotypes (Panel A of Figure 1). The mean posterior probabilities of membership of the latent class were 0.841, 0.802 and 0.874, respectively, which were large and therefore showed that the classes were well-separated. Class 1 (46.1%) corresponds to patients with a milder presentation of MIS-C not requiring intensive care, with no or minimal cardiac involvement. Class 2 (25.3%) represents children with a severe presentation of MIS-C, with cardiac system involvement and < 4 organ systems involved, and Class 3 (28.6%) represents children with the more severe presentation of MIS-C including cardiac involvement along with > 4 systems involved, including respiratory, gastrointestinal (GI), renal, hematologic, and dermatological manifestations. These findings suggest that children with GI presentations or skin rashes may have an increased risk of more severe disease. Besides, dMLCA estimated the site-specific prevalence of subphenotypes (Panel B of Figure 1) to help understanding the population composition in each site. Reasons to explain the heterogeneity across sites include differences in race/ethnicity distributions, patient acuity, and evaluation and treatment protocols for children with MIS-C.

To demonstrate the validity of the study design, we showed that the MIS-C latent classes we found were unique from the general COVID-19 cohort. The heatmap of the characteristics of latent classes of COVID-19 PCR-positive children without MIS-C diagnoses was very different from the characteristics of MIS-C latent classes (Panel A of Figure 2). We further visualized the distances among COVID-19 and MIS-C subpopulations (Panel B of Figure 2). The latent classes of the two cohorts were separated from each other in general except for a closeness among MIS-C Classes 1, 2, and COVID-19 Class 1. This indicated the clinical overlap in some of the presentations of MIS-C and acute COVID-19, which has been described in earlier studies of MIS-C [21].

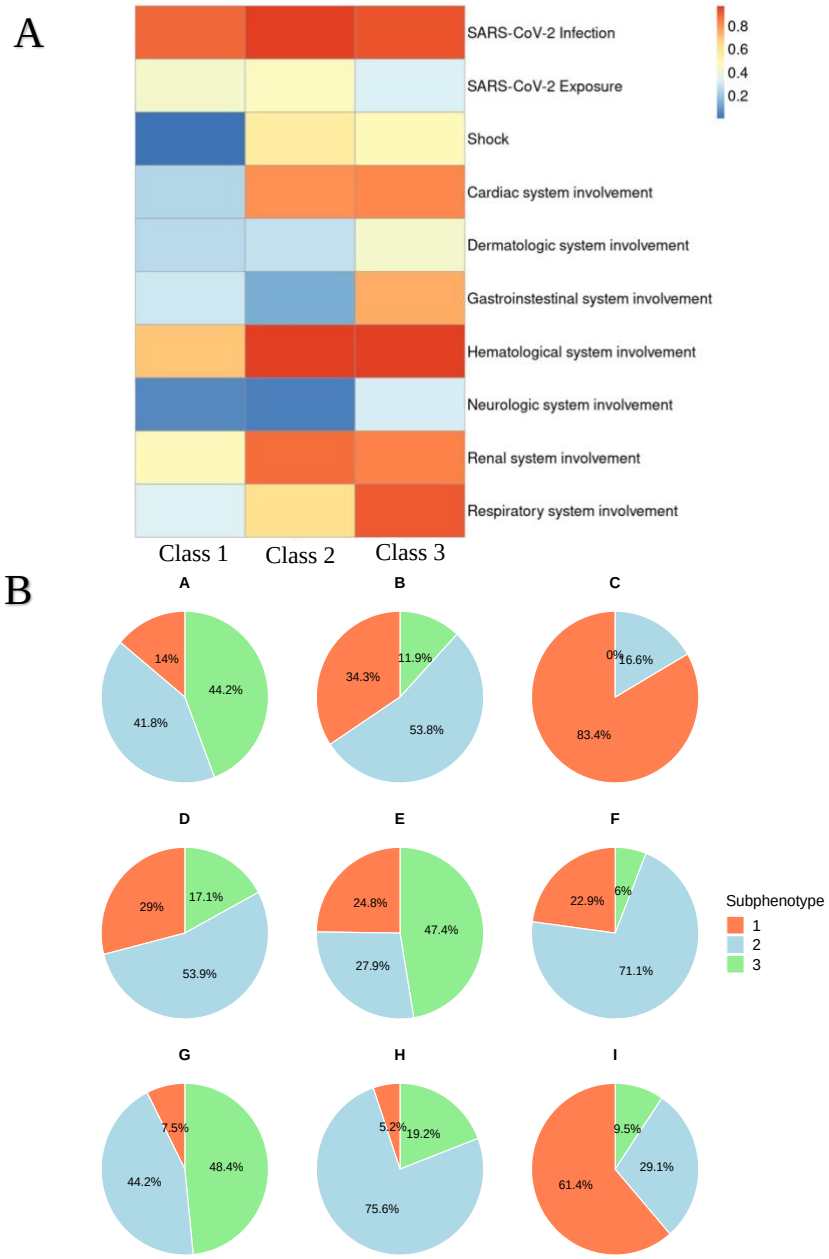

Figure 1 Results of MIS-C data analysis using dMLCA with three latent classes. A) Heatmap showing the prevalence of ten manifest variables in three latent classes; Each column represents a latent class, and each row represents a manifest variable. The color of the boxes represents the prevalence. The legend on the top right shows the scale of the colors. Red represents prevalence close to 100% and blue represents prevalence close to 0%. Class 1 corresponds to patients with a milder presentation of MIS-C not requiring intensive care, with no or minimal cardiac involvement. Class 2 represents children with a severe presentation of MIS-C, with cardiac system involvement and < 4 organ systems involved, and Class 3 represents children with the more severe presentation of MIS-C including cardiac involvement along with ≥ 4 systems involved, including respiratory, gastrointestinal (GI), renal, hematologic, and dermatological manifestations. **B)** Pie charts showing the prevalence of the three latent classes by site.

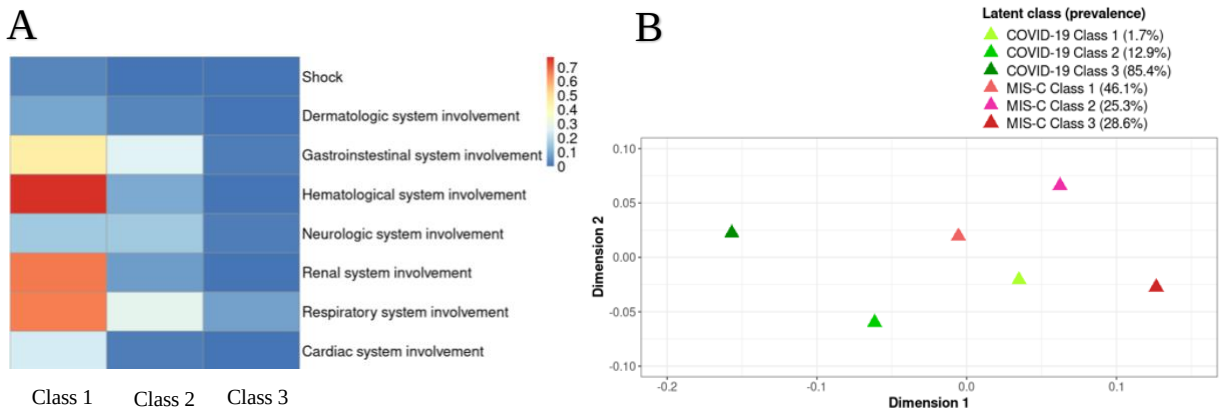

Figure 2. A) Heatmap showing the latent classes and their characteristics of children testing positive for SARS-CoV-2 by PCR test. **B)** 2-Dimensional plot comparing the distances among MIS-C and COVID-19 PCR positive subpopulations. Closer subpopulations have larger similarities. The distance between each pair of latent classes was measured by fixation index (Fst) and mapped onto a 2-dimensional plot through multidimensional scaling.

Conclusion

The dMLCA algorithm is an effective federated learning algorithm for disease subphenotyping under a distributed environment. It is the first distributed-EM algorithm for learning subphenotypes that can be applied for consistent clustering across multiple institutions. It can be directly applied to multi-center OHDSI studies for clustering tasks as long as the data are harmonized by a CDM (e.g., OMOP CDM). We've compared the local performance versus multi-site performance through simulation (not included in this report) and showed that the multi-site algorithm provides more accurate estimation. Our method advances the methodology development in federated unsupervised learning and contributes to generating reproducible and reliable evidence using real-world data to answer clinical questions.

A limitation is that multiple communication rounds among institutions are needed to achieve the optimal result. We are now working on making the iterative communications be few-shots and developing an R package based on this version.

Reference

1. Chen Y, Dong G, Han J, et al. Regression cubes with lossless compression and aggregation. *IEEE Trans Knowl Data Eng* 2006;18(12):1585–99.
2. Luo, C., Islam, M.N., Sheils, N.E. *et al.* DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat Commun* **13,** 1678 (2022).
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;19(5):758–64.
4. Duan R, Boland MR, Moore JH, et al. ODAL: A One-Shot Distributed Algorithm to Perform Logistic Regressions on Electronic Health Records Data from Multiple Clinical Sites. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium* 2018;30-41.

5. Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *J Am Med Inform Assoc* 2020;27(3):376-85.
6. Lu C-L, Wang S, Ji Z, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc* 2015;22(6):1212–9.
7. Duan, Rui, et al. "Learning from local to global: An efficient distributed algorithm for modeling time-to-event data." *Journal of the American Medical Informatics Association* 27.7 (2020): 1028-1036.
8. Luo, C., Duan, R., Naj, AC et al. ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. *Sci Rep* 2022;12(1):6627.
9. Edmondson, M.J., Luo, C., Duan, R. et al. An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes. Sci Rep. 2021;11(1):19647.
10. Zhang Y, Duchi JC, Wainwright MJ. Communication-efficient algorithms for statistical optimization. *J Mach Learn Res* 2013;14(1):3321-63.
11. Lee JD, Liu Q, Sun Y, et al. Communication-efficient sparse regression. *J Mach Learn Res* 2017;18(1):115-44.
12. Battey H, Fan J, Liu H, et al. Distributed testing and estimation under sparse high dimensional models. *Ann Stat* 2018;46(3):1352.
13. Dobriban E, Sheng Y. Distributed linear regression by averaging. *Ann Stat* 2021;49(2):918-43.
14. Dobriban E, Sheng Y. WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions. *J Mach Learn Res* 2020;21(66):1-52.
15. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. *J Am Stat Assoc* 2018;114:668–81.
16. Wang J, Kolar M, Srebro N, et al. Efficient distributed learning with sparsity. In proceedings of the 34th International Conference on Machine Learning 2017;70:3636-45.
17. Teng C, Thampy U, Bae JY, et al. Identification of Phenotypes Among COVID-19 Patients in the United States Using Latent Class Analysis. Infect Drug Resist. 2021;14:3865-3871.
18. Calfee CS, Delucchi K, Parsons PE, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. Lancet Respir Med. 2014;2(8):611-620.
19. Cai T, Liu M, Xia Y. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. J Am Stat Assoc 2021;1-34.
20. Duan R, Ning Y, Chen Y. Heterogeneity-aware and communication efficient distributed statistical inference. Biometrika 2021 (in press).
21. Godfred-Cato S, Bryant B, Leung J, et al. COVID-19-Associated Multisystem Inflammatory Syndrome in Children - United States, March-July 2020. MMWR Morb Mortal Wkly Rep. 2020;69(32):1074-1080.