Journal of the American Medical Informatics Association, 29(8), 2022, 1366–1371
https://doi.org/10.1093/jamia/ocac067
Advance Access Publication Date: 17 May 2022
Research and Applications

Research and Applications

dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling

Chongliang Luo[1,2], Md. Nazmul Islam[3], Natalie E. Sheils[3], John Buresh[3], Martijn J. Schuemie[4], Jalpa A. Doshi[5,6], Rachel M. Werner[5,6,7], David A. Asch[5,6], and Yong Chen [1,6]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [2]Division of Public Health Sciences, Washington University School of Medicine in St. Louis, St Louis, Missouri, USA, [3]OptumLabs, Minnetonka, Minnesota, USA, [4]Janssen Research and Development LLC, Titusville, New Jersey, USA, [5]Division of General Internal Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [6]Leonard Davis Institute of Health Economics, Philadelphia, Pennsylvania, USA, and [7]CpI Michael J Crescenz VA Medical Center, Philadelphia, Pennsylvania, USA

Corresponding Author: Yong Chen, PhD, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 602 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA; ychen123@upenn.edu

Received 12 January 2022; Revised 19 April 2022; Editorial Decision 21 April 2022; Accepted 10 May 2022

ABSTRACT

Objective: To develop a lossless distributed algorithm for generalized linear mixed model (GLMM) with application to privacy-preserving hospital profiling.

Materials and Methods: The GLMM is often fitted to implement hospital profiling, using clinical or administrative claims data. Due to individual patient data (IPD) privacy regulations and the computational complexity of GLMM, a distributed algorithm for hospital profiling is needed. We develop a novel distributed penalized quasi-likelihood (dPQL) algorithm to fit GLMM when only aggregated data, rather than IPD, can be shared across hospitals. We also show that the standardized mortality rates, which are often reported as the results of hospital profiling, can also be calculated distributively without sharing IPD. We demonstrate the applicability of the proposed dPQL algorithm by ranking 929 hospitals for coronavirus disease 2019 (COVID-19) mortality or referral to hospice that have been previously studied.

Results: The proposed dPQL algorithm is mathematically proven to be lossless, that is, it obtains identical results as if IPD were pooled from all hospitals. In the example of hospital profiling regarding COVID-19 mortality, the dPQL algorithm reached convergence with only 5 iterations, and the estimation of fixed effects, random effects, and mortality rates were identical to that of the PQL from pooled data.

Conclusion: The dPQL algorithm is lossless, privacy-preserving and fast-converging for fitting GLMM. It provides an extremely suitable and convenient distributed approach for hospital profiling.

Key words: distributed penalized quasi-likelihood algorithm, federated learning, generalized linear mixed model, hospital profiling, privacy-preserving

INTRODUCTION

Decades of health services research have revealed that the outcomes hospitalized patients achieve are considerably determined by where they are admitted. Hospital profiling allows a quantitative assessment of the quality of hospital care that may help patients decide which hospital to use, or may guide how those hospitals are accredited or paid. Studying cross-hospital variation in care helps identify reasons for that variation with the aim of improving care for all. Such profiling across hospitals is usually conducted by analyzing clinical or administrative insurance claims data, always considering what factors to adjust for statistically—for example, patient characteristics like sociodemographic or medical conditions, hospital characteristics like volume or academic status, or social or community characteristics like area-level poverty or uninsurance levels.[1,2] In a recent article on hospital profiling for coronavirus disease 2019 (COVID-19) mortality,[2] Asch et al. ranked the performance of 929 hospitals after adjusting for the patients' characteristics including age, sex, Elixhauser comorbidities, insurance type, and hospital's characteristics including number of beds, number of intensive care unit beds, urban/nonurban setting, geographic region, profit status, and academic affiliation. Research of this kind helps untangle what are often separate contributors to the production of good patient outcomes and is essential for identifying ways to improve those outcomes.

Recent years have seen the development of statistical methodologies for the purpose of hospital profiling. A commonly used model is the generalized linear mixed model (GLMM), which assumes common fixed-effects of covariates, for example, patient- and hospital-level factors, and hospital-specific random effects, that is, intercepts on the interested clinical outcome.[1–3] Based on the estimated fixed and random effects, the risk standardized event rates (RSERs) can be calculated for each site. GLMM estimation, though complicated, could be obtained by methods such as Gaussian-Hermite approximation of the integrated likelihood, Monte-Carlo-based approaches, and the penalized quasi-likelihood (PQL) approach.[4,5]

For example, Drye et al,[6] studied the in-hospital and 30-day mortality rate of acute myocardial infarction (AMI), heart failure (HF), and pneumonia for more than 3000 hospitals using Medicare claims data from the Centers for Medicare and Medicaid Services (CMS). Asch et al[2] studied COVID-19 mortality or discharge to hospice in 929 hospitals using the UnitedHealth Group Clinical Discovery Portal. Both investigations were based on a large integrated database, where patient-level data from multiple hospitals were available in a single dataset. But often such integrated datasets are not available. Indeed, an important limitation of most other investigations is that they rely on data sets from single institutions and so are smaller, more homogenous, and less representative of broader populations.

Ideally, if individual patient-level datasets from across multiple payers and institutes could be shared, the profiling methods can be applied to a larger and more general study population. However, it is often the case that these individual patient-level data are typically protected by privacy regulations and sharing of individual patient data (IPD) is difficult. To extend hospital profiling to cover a wider spectrum of patient populations, privacy-preserving distributed algorithms can be used. Specifically, when fitting GLMM, the distributed algorithm is expected to require aggregated data (AD) from each hospital (often iteratively) but obtains accurate estimates of the model parameters, and therefore accurate estimates of RSERs. Recently, Zhu et al[7] proposed a distributed algorithm based on Expectation–Maximization (EM) that involves the Metropolis-Hasting algorithm. However, it is well known that the EM algorithm usually takes many iterations to converge—the distributed EM algorithm of Zhu et al[7] requires 500∼1000 iterations for results to be converged. As a result, the distributed algorithm also requires many rounds of data communication between institutes.

This article aims to fill this important methodological gap by proposing a novel distributed algorithm to fit GLMM that is lossless (ie, it obtains identical results as if the IPD are pooled from all hospitals), computationally stable, and, importantly, *requires only a few rounds of communications of AD across institutes*. The algorithm is based on the PQL approach and a newly developed distributed algorithm for linear mixed model (LMM). We demonstrate the applicability of the proposed distributed PQL (dPQL) algorithm by hospital profiling for COVID-19 mortality or referral to hospice using data from 929 hospitals that have been previously studied by Asch et al.[2]

METHODS

Fitting GLMM via penalized quasi-likelihood

GLMM is an extension of GLM with random effects. We introduce notations of GLMM in the context of hospital profiling. Assume there are K hospitals with numbers of patients n_i, and the total number of patients is $N = \sum_i n_i$. For subject j at hospital i, we denote y_{ij} the outcome, x_{ij} the p-dimensional covariates with fixed effects β, and u_i the random effect (ie, random intercept), $i = 1, \ldots, K,\ j = 1, \ldots, n_i$. Conditional on the covariates $X_i = (x_{i1}, \ldots, x_{in_i})^T$ and random effects u_i, $y_i = (y_{i1}, \ldots, y_{in_i})^T$ are assumed to be independent observations with means and variances specified by a GLM. Specifically,

$$E(y_{ij}|u_i) = \mu_{ij} = h(\eta_{ij}) = h(x_{ij}{}^T\beta + u_i), \quad (1)$$
$$Var(y_{ij}|u_i) = v(\mu_{ij}), \quad (2)$$

where $g = h^{-1}$ is the link function that connects the conditional means μ_{ij} to the linear predictor η_{ij}, $v(\cdot)$ is the variance function. The random effects u_i are assumed to follow a normal distribution with mean 0 and variance θ. We note that the above model dictated by Equations (1) and (2) could be extended to hierarchical models as in George et al[3] for more flexibilities; for example, the covariate X_i can include hospital-level characteristics (eg, the (log) volume of a hospital) and the variance of the random effects u_i could also be dependent on the hospital-level characteristics.

Standard estimation of the GLMM parameters (β, θ) is based on maximizing the integrated quasi-likelihood

$$L(\beta, \theta) = \{2\pi\theta\}^{-K/2} \prod_{i=1}^{K} \int_{-\infty}^{\infty} \exp[-\sum_{j=1}^{n_i} d_{ij}(y_{ij}, \mu_{ij})/2 - u_i{}^T\theta^{-1}u_i/2]du_i,$$

where

$$d_{ij}\left(y, \mu\right) = -2\int_{y}^{\mu}(y-u)/v(u)\ du.$$

Maximization of this objective function is generally complicated,[4] as the integrations must be performed numerically unless in the case of Gaussian outcome and identity link.

One approach to the integration is to make a Laplace approximation, which eventually leads to the PQL algorithm.[4] The PQL algorithm iteratively fit the linear mixed model

$$y_{ij}{}^* = x_{ij}{}^T\beta + u_i + \epsilon_{ij},\ \ \epsilon_{ij} \sim N(0, w_{ij}{}^{-1}), \quad (3)$$

with the working outcome

The proposed dPQL algorithm

1. Initialize: the lead site send an initial value of the fixed effects $\beta^{(0)}$, and the random effects $u_i^{(0)} = 0$ to the collaborative sites $i = 1, \ldots, K$.
2. For iteration $s = 0, 1, \ldots,$
   2.1 Site i calculates the working outcome $y_i^* = \eta_i^{(s)} + (y_{ij} - \mu_i^{(s)})g'(\mu_i^{(s)})$, $\eta_i^{(s)} = X_i\beta^{(s)} + u_i^{(s)}$, and the weights $W_i = diag\{g'(\mu_i^{(s)})^{-2}v(\mu_i^{(s)})\}$,
   2.2 Site i calculates aggregated data
   - $p \times p$ matrix: $S_i^X = X_i^T W_i X_i$,
   - $p-$ dim vector: $S_i^{Xy} = X_i^T W_i y_i^*$ and
   - scalars: $s_i^y = y_i^{*T} W_i y_i^*$ and sample size n_i, and transfers them to the lead site,

   2.3 The lead site fits weighted DLMM algorithm based on the aggregated data from 2.2, to obtain updated $\beta^{(s+1)}$, $u_i^{(s+1)}$, and send them back to the collaborative sites.
3. Stop iteration when converged, for example, $||\eta^{(s+1)} - \eta^{(s)}||/||\eta^{(s)}|| < 1e\text{-}6$. The final estimates are $\hat{\beta} = \beta^{(s)}$, $\tilde{u}_i = u_i^{(s)}$ and $\hat{\theta}$.

$$y_{ij}^* = x_{ij}^T\hat{\beta} + \tilde{u}_i + (y_{ij} - \hat{\mu}_{ij})g'(\hat{\mu}_{ij}), \tag{4}$$

and the weight

$$w_{ij} = g'(\hat{\mu}_{ij})^{-2}v(\hat{\mu}_{ij}). \tag{5}$$

The obtained estimates are denoted as $(\hat{\beta}, \hat{\theta})$. See Breslow et al[4] for more details about the derivation.

The proposed dPQL algorithm

We develop a dPQL algorithm for GLMM estimation in the case that the IPD are distributed across multiple centers and direct transfer of the IPD is not allowed. The dPQL algorithm is based on the distributed linear mixed model (DLMM) algorithm, which fits LMM exactly by requiring each site to contribute some AD only once.[8] Specifically, in each iteration of the PQL algorithm, the weighted LMM (3) is fitted by the DLMM algorithm, requiring each site to contribute AD

- $p \times p$ matrix $S_i^X = X_i^T W_i X_i$,
- $p - dim$ vector $S_i^{Xy} = X_i^T W_i y_i^*$, and
- scalars $s_i^y = y_i^{*T} W_i y_i^*$, and sample size n_i.

See the Supplementary Materials for details of the DLMM algorithm. The dPQL algorithm thus reconstructs the PQL iterations and obtains identical results as if the IPD are pooled together.

Distributed calculation for standardized mortality rates based on dPQL

Hospital profiling results are often reported with the standardized mortality rates (SMRs) of hospitals. We demonstrate that the SMRs of hospitals can also be calculated in a privacy-preserving fashion. We provide 2 approaches for risk standardization, the Indirectly Standardized Mortality Rate (denoted as ISMR)[1] and the Directly Standardized Mortality Rates (denoted as DSMR).[3,6] While both approaches measure adjusted mortality rates effectively, DSMR in contrast to SMR, has an interpretation in an amenable probability scale.

The ISMR of hospital k is estimated[1] as

$$\text{ISMR}_k = \frac{\hat{p}_k}{\hat{e}_k} \times \bar{y}, \tag{6}$$

where

$$\hat{p}_k = n_k^{-1}\sum_{j=1}^{n_k} h(x_{kj}^T\hat{\beta} + \tilde{u}_k) \tag{7}$$

is the average expected mortality rate for patients at hospital k,

$$\hat{e}_k = n_k^{-1}\sum_{j=1}^{n_k} h(x_{kj}^T\hat{\beta}) \tag{8}$$

is the average expected mortality rate for hospital k patients had they been treated at the "population level," and $\bar{y}$ is the overall observed mortality rate. This SMR measure has been used to compare the performance of nonfederal acute care hospitals in the United States for AMI ($n = 3135$ hospitals), HF ($n = 4209$ hospitals), and pneumonia ($n = 4498$ hospitals) from 2004 to 2006.[6]

The DSMR of hospital k is defined as the average mortality rate assuming patients from all the hospitals being treated at this hospital,[4] thai is,

$$\text{DSMR}_k = N^{-1}\sum_{i=1}^{K} n_i\hat{p}_{ik}, \tag{9}$$

where

$$\hat{p}_{ik} = n_i^{-1}\sum_{j=1}^{n_i} h(x_{ij}^T\hat{\beta} + \tilde{u}_k) \tag{10}$$

is the average expected mortality rate of patients at hospital i had they been treated at hospital k. When $i = k$, $\hat{p}_{ik} = \hat{p}_k$, and if $i \neq k$, $\hat{p}_{ik}$ is a counterfactual probability. This SMR measure has been applied to profiling 4289 hospitals in the United States for AMI using Medicare records from 2009 to 2011,[3] and to evaluating COVID-19 mortality in 929 hospitals.[2] While both approaches measure adjusted mortality rates effectively, DSMR in contrast to ISMR, has an interpretation in an amenable probability scale.

We note that both types of SMR measures (ISMR and DSMR) can be calculated distributively without sharing IPD. Specifically, for the ISMR, each individual hospital calculates and shares its average expected mortality rates (ie, 2 probabilities $\hat{p}_{kk}$ and $\hat{e}_k$ as in Figure 1, and Equations 7 and 8) using its own patient-level data and the public estimates from dPQL (ie, $\hat{\beta}$ and $\tilde{u}$ as in Figure 1). For the DSMR, each individual hospital calculates and shares the average expected mortality rates had its patients been treated at other hospitals (ie, K probabilities $\hat{p}_{i1}, \ldots, \hat{p}_{iK}$ as in Figure 1, and Equation 10) using its own patient-level data and the public estimates from dPQL.

Us hospital ranking based on the mortality rates for patients admitted with COVID-19

Asch et al[2] conducted a cohort study assessing 38 517 adults who were admitted with COVID-19 to 929 US hospitals from

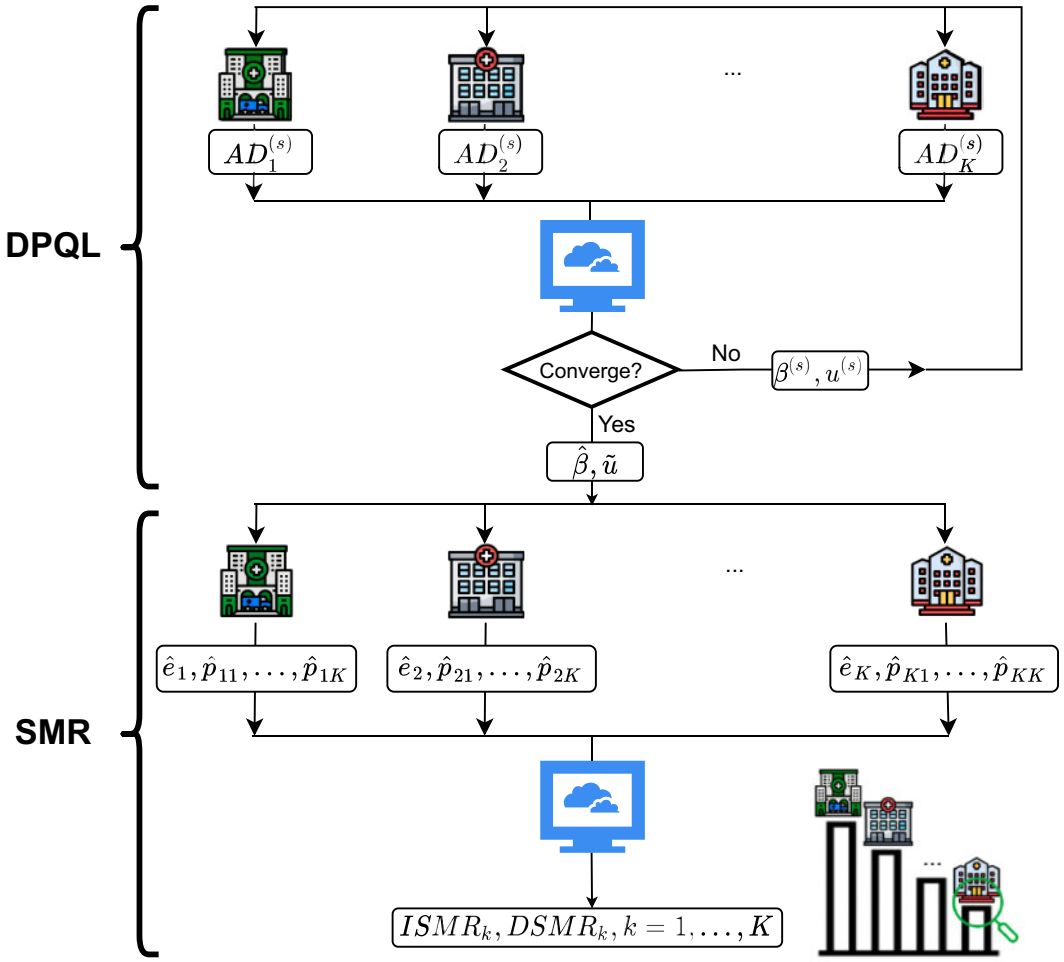

Figure 1. A distributed procedure for hospital profiling. The dPQL algorithm fit the GLMM in a distributive fashion by requiring some aggregated data (AD) from each hospital in a few iterations, and obtains the estimated fixed effects ($\widehat{\beta}$) and random effects ($\tilde{u}$). Next, standardized mortality rates (SMRs) of the hospitals can be calculated distributively. Based on the results of dPQL algorithm, each hospital calculates its average expected mortality rates using its own individual patient data (ie, for hospital k, $\hat{e}_k$ is the average expected mortality rate had its patients been treated at the "population level," and $\hat{p}_{ki}$ is the average expected mortality rates had its patients been treated at hospital $i = 1, \ldots, K$). The indirectly and directly standardized mortality rates can then be calculated ($ISMR_k$ and $DSMR_k$ for hospital k).

January 1, 2020 to June 30, 2020 using the data from United-Health Group Clinical Discovery Portal. The hospital's standardized rate of 30-day in-hospital mortality or referral to hospice was calculated, after adjusting for patient-level characteristics, including demographic data, Elixhauser comorbidities,[9] community or nursing facility admission source, and time since January 1, 2020; hospital-level characteristics, including size, the number of intensive care unit beds, academic and profit status, hospital setting; and regional characteristics, including COVID-19 case burden. See Supplementary Figure S1 for a description of the data.

We demonstrate the applicability of the proposed dPQL algorithm by using it to rank hospitals with only transferring AD from each hospital. Specifically, we compare the predicted mortality rate (via ISMR or DSMR) of the 929 hospitals by either pooled analysis (PQL) of the patient-level data or the distributed analysis (dPQL) of the AD across hospitals. We also check the number of iterations for reaching convergence, and compare the estimation of fixed effects, best linear unbiased predictors (BLUPs), and mortality rates using either pooled or distributed analyses.

RESULTS

The predicted mortality rate (via ISMR or DSMR) of the 929 hospitals by either pooled analysis (PQL) or the distributed analysis (dPQL) is compared in Figure 2. The dPQL algorithm reached convergence with only 5 iterations, and the estimation of fixed effects, BLUPs, and mortality rates were identical to that of the PQL from pooled data. The estimated fixed and random effects from the dPQL algorithm and from the PQL are also identical, as shown in Supplementary Figure S2.

DISCUSSION

We propose a novel dPQL algorithm, a privacy-preserving distributed learning algorithm to fit GLMM. The dPQL algorithm does not require sharing of individual patient-level data. The algorithm only requires sharing of minimal AD from each site over few rounds of communication and obtains identical results as if fitting GLMM to the pooled data using PQL. The calculation of AD at each individual site is implemented in the R package "pda."[10,11] We also de-

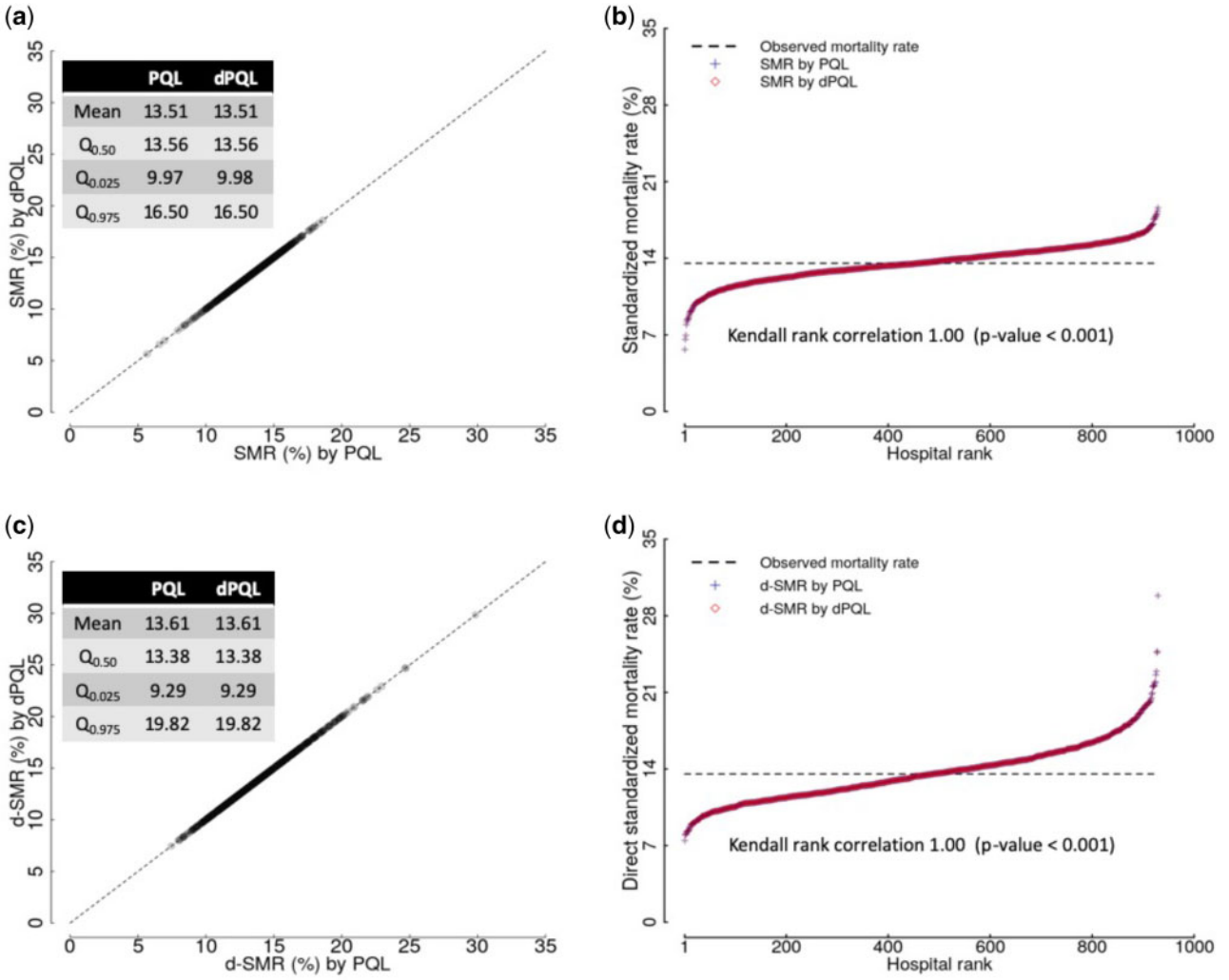

Figure 2. The estimated mortality rate (indirectly standardized mortality rate (A) and (B) and directly standardized mortality rates (C) and (D)) of 30-day in-hospital mortality or referral to hospice of the 929 hospitals by either pooled analysis (PQL) or the distributed analysis (dPQL).

veloped an "over-the-air" online portal called PDA-OTA (http://pda-ota.pdamethods.org/) to facilitate secure and convenient collaboration on the basis of the "pda" package. See the Supplementary Materials for detailed instructions for using the PDA-OTA.

The results of the PQL estimation are comparable to that of other approaches used to fit the GLMM model. For example, in the hospital ranking for COVID-19 mortality rates, the PQL estimation is almost identical to that of the Gaussian-Hermite approximation approach used in the original paper.[2] Although fitting GLMM by PQL is sometimes criticized for its biased estimation when the outcome is binary and clusters are small,[4,5] it is still an appropriate estimation approach for hospital profiling purposes, as the sample sizes in hospitals are usually large enough.

The communication efficiency of the dPQL algorithm is attributable to the fast convergence of the PQL algorithm. See Supplementary Figure S3 showing the convergence in just a few iterations in the hospital profiling example for COVID-19 mortality. The communication efficiency can be further improved by a one-shot (or few-shots) version of the dPQL algorithm, that is, run only one (or few) iteration of the dPQL algorithm proposed in Section "The proposed dPQL algorithm." Such a one-shot approach has been pursued by many distributed algorithms and is considered communication-efficient.[8,12–17] The one-shot version dPQL algorithm will sacrifice some accuracy of the estimation, but obtains very appealing communication cost, as each hospital needs only to share the AD once. Meanwhile, the number of iterations required in the PQL algorithm depends on the choice of initial values. While default initial values (ie, all fixed effects being 0) provide satisfactory results, the performances can be improved with smart choices of initial values. We recommend setting a maximum number of iterations (eg, within 5 iterations) when using dPQL in practice. However, we do not recommend applying dPQL in the high-dimensional setting (ie, large p) as it will involve communication of massive aggregate data (ie, the p-by-p matrices S_i^X).

We provide indirect (ISMR) and direct (DSMR) standardization to interpreting the hospital ranking for the purpose of public reporting. Examples of conducting hospital profiling using either approach exist in literature.[2,3,6] The directly standardized approach is considered to behave better for models that consider the interaction between the hospital and the patients.[4] On the other hand, using GLMM for ranking hospitals assumes overlap of patient characteristics at different hospitals. Other statistical models, for example, without random effects, could also be considered when there is poor overlap of patient characteristics between hospitals. The choice of standardization approaches and statistical models is beyond the scope of this paper. The hospital

profiling can be conducted for other tasks, as long as the outcome can be modeled by GLMM. This includes binary outcomes such as COVID-related mortality, ventilator usage or hospital readmission, and count outcomes such as hospitalization length of stay, etc.

Our proposed dPQL algorithm is in a similar fashion as federated learning methods, which have found profound applications in many clinical settings in recent years.[18] However, our AD release mechanism has not been investigated in rigorous privacy framework such as k-anonymity[19] or differential privacy,[20,21] and thus is not guaranteed to be protected from the risk of re-identification or membership inference attacks[22] (MIAs). Specifically, the risk of re-identification arises from linking potential quasi-identifiers (eg, combinations of patient's characteristics) to external sources,[19] and the risk of MIAs refers to inferring whether a data point (eg, a specific patient's record) is used to train the model.[22,23] To avoid potential risk of re-identification, we suggest excluding or suppressing values representing 10 or fewer patients[24] when sharing the aggregate data and using random initial values if possible when initiating the iteration. These will prevent the aggregate data from containing sparse elements and hence re-identifying sensitive patient information.[8] We also suggest avoiding high-dimensional GLMM models, and using a representative sample for training. These will prevent overfitting and improve the generalizability of the model, which result in mitigating the risk of MIAs.[22] In the future, we plan to extend our dPQL algorithm via techniques such as differential privacy and multiparty homomorphic encryption.[25]

FUNDING

This work was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2019C3-18315). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

AUTHOR CONTRIBUTIONS

CL and YC designed methods and analyses; MNI, NES, and JB provided the dataset from UnitedHealth Group; CL and MNI conducted numerical analyses; all authors interpreted the results and provided instructive comments; CL and YC drafted the main article. All authors have approved the article.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

NES and MNI and JB are or were full-time employees in Optum Labs and own stock in its parent company, UnitedHealth Group, Inc. The other authors have no competing interests to declare.

DATA AVAILABILITY STATEMENT

All data were accessed in compliance with the HIPPA rules, IRB approval or waiver of authorization was not required.

REFERENCES

1. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Stat. Sci* 2007; 22: 206–26.
2. Asch DA, Sheils NE, Islam MN, *et al.* Variation in US hospital mortality rates for patients admitted with COVID-19 during the first 6 months of the pandemic. *JAMA Intern Med* 2021; 181 (4): 471–8.
3. George EI, Ročková V, Rosenbaum PR, Satopää VA, Silber JH. Mortality rate estimation and standardization for public reporting: Medicare's hospital compare. *J Am Stat Assoc* 2017; 112 (519): 933–47.
4. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993; 88 (421): 9–25.
5. Breslow N. Whither PQL? In: *Proceedings of the Second Seattle Symposium in Biostatistics*. Springer, 2004: 1–22; New York, NY.
6. Drye EE, Normand S-LT, Wang Y, *et al.* Comparison of hospital risk-standardized mortality rates calculated by using in-hospital and 30-day models: an observational study with implications for hospital profiling. *Ann Intern Med* 2012; 156 (1 Pt 1): 19–26.
7. Zhu R, Jiang C, Wang X, *et al.* Privacy-preserving construction of generalized linear mixed model for biomedical computation. *Bioinformatics* 2020; 36 (Suppl_1): i128–i135.
8. Luo C, *et al.* DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat Commun* 2022; 13 (1678): 1–10.
9. Thompson NR, Fan Y, Dalton JE, *et al.* A new Elixhauser-based comorbidity summary measure to predict in-hospital mortality. *Med Care* 2015; 53 (4): 374–9.
10. Luo C, Duan R, Edmondson M, Tong J, Chen Y. PDA: Privacy-Preserving Distributed Algorithms. *R Package 1.0-2*. 2020-12-10. https://CRAN.R-project.org/package=pda.
11. Luo C, Duan R, Edmondson M, Jiayi T, Chen Y. PDA: Privacy-Preserving Distributed Algorithms (v 1.2-1). *Github* 2021-12-01. https://github.com/Penncil/pda.
12. Battey H, Fan J, Liu H, Lu J, Zhu Z. Distributed testing and estimation under sparse high dimensional models. *Ann Stat* 2018; 46 (3): 1352–82.
13. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. *J Am Stat Assoc* 2019; 114 (526): 668–81.
14. Duan R, Luo C, Schuemie MJ, *et al.* Learning from local to global-an efficient distributed algorithm for modeling time-to-event data. *J Am Med Inform Assoc* 2020; 27 (7): 1028–36.
15. 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.
16. Luo C, Duan R, Naj AC, Kranzler HR, Bian J, Chen Y. ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. *Sci Rep* 2022; 12 (1): 6627.
17. Duan R, Ning Y, Chen Y. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika* 2022; 109 (1): 67–83.
18. Rieke N, Hancox J, Li W, *et al.* The future of digital health with federated learning. *NPJ Digit Med* 2020; 3: 1–7.
19. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Unc Fuzz Knowl Based Syst* 2002; 10 (05): 557–70.
20. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. *J Priv Confidentiality* 2017; 7 (3): 17–51.
21. Wasserman L, Zhou S. A statistical framework for differential privacy. *J Am Stat Assoc* 2010; 105 (489): 375–89.
22. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017: 3–18.
23. Pyrgelis A, Troncoso C, De Cristofaro E. Knock knock, who's there? Membership inference on aggregate location data. *arXiv Prepr. arXiv1708.06145* 2017.
24. CMS Cell Suppression Policy. https://www.hhs.gov/guidance/document/cms-cell-suppression-policy Accessed April 15, 2022.
25. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, *et al.* Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun 2021; 12 (1): 5910.