JOHNS HOPKINS
UNIVERSITY

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01846-1

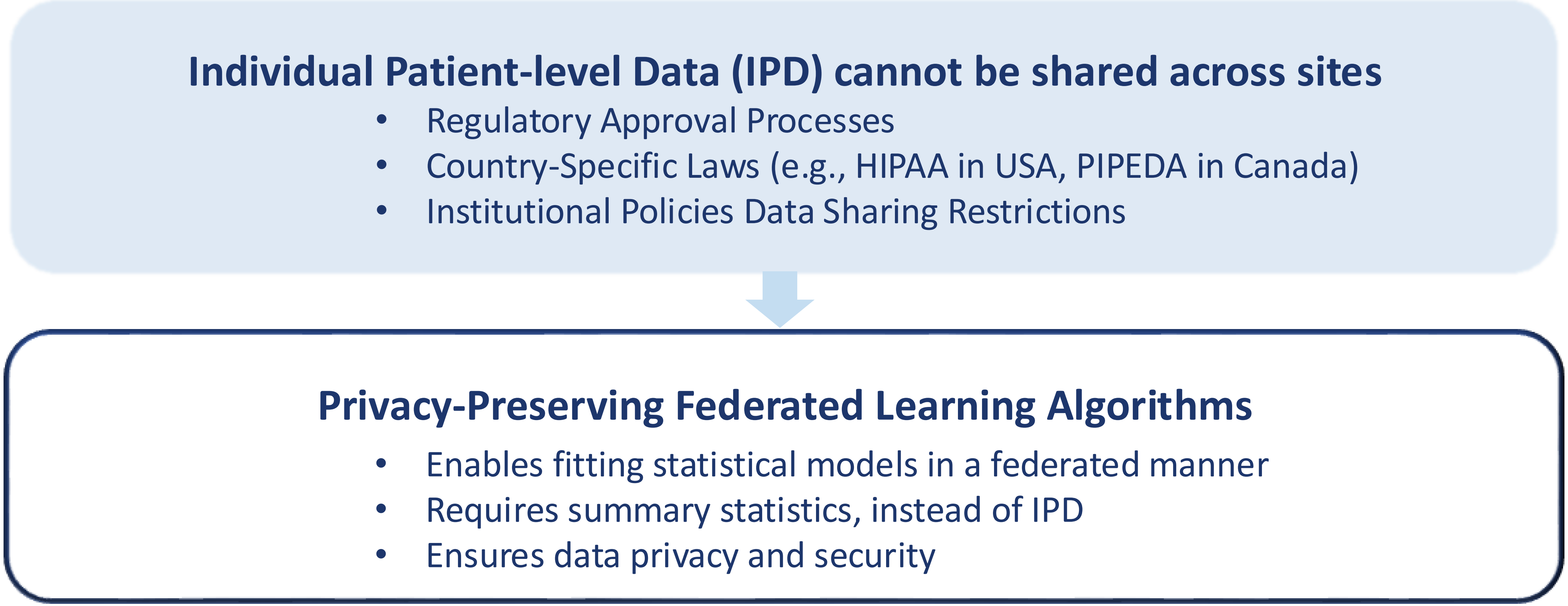
Unlocking efficiency in real-world collaborative studies: a multi-site international study with one-shot lossless GLMM algorithm

Check for updates

Jiayi Tong[1,2,3] ✉, Jenna M. Reps[4,5,6], Chongliang Luo[7], Yiwen Lu[1,2], Lu Li[1,2], Juan Manuel Ramirez-Anguita[8], Milou T. Brand[9], Scott L. DuVall[10,11], Thomas Falconer[12], Alex Mayer Fuentes[13], Xing He[14,15], Michael E. Matheny[16,17], Miguel A. Mayer[8], Bhavnisha K. Patel[16,17], Katherine R. Simon[16,17], Marc A. Suchard[11,18], Guojun Tang[19], Benjamin Viernes[11], Ross D. Williams[6], Mui van Zandt[9], Fei Wang[20], Jiang Bian[14,15], Jiayu Zhou[21], David A. Asch[22,23] & Yong Chen[1,2,23] ✉

Jiayi (Jessie) Tong
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
September 16, 2025, OHDSI Community-Call

A **Primary Challenge** in Multi-site International Study



Individual Patient-level Data (IPD) cannot be shared across sites

- Regulatory Approval Processes
- Country-Specific Laws (e.g., HIPAA in USA, PIPEDA in Canada)
- Institutional Policies Data Sharing Restrictions

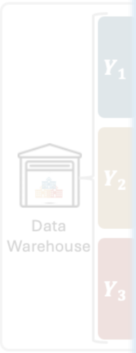
Privacy-Preserving Federated Learning Algorithms

- Enables fitting statistical models in a federated manner
- Requires summary statistics, instead of IPD
- Ensures data privacy and security

Two Ideal Properties of Federated Learning Algorithms

Lossless

One-shot



Pool

Results

To date, only a few algorithms have successfully achieved both lossless and one-shot properties simultaneously:

- Linear Regression (i.e., Chen et al., 2006, IEEE)
- Linear mixed models (i.e., Luo et al., 2022, Nature Communications)

Identical Results

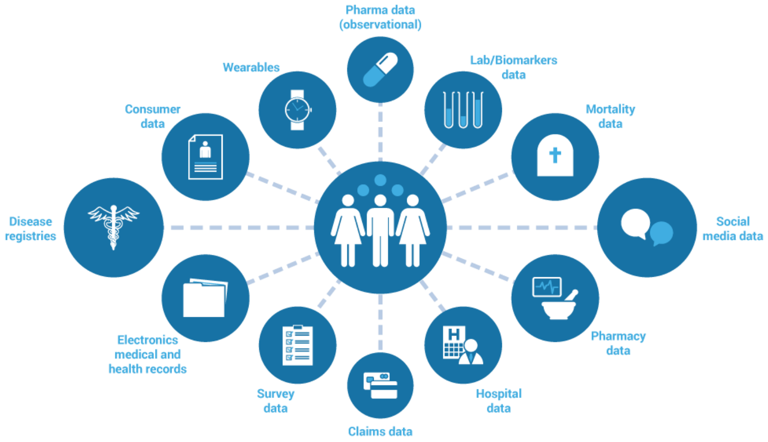
No accuracy loss due to data sharing constraints

Only a single round of communication is needed

Luo C, Islam MN, Sheils NE, Buresh J, Reps J, Schuemie MJ, Ryan PB, Edmondson M, Duan R, Tong J, Marks-Anglin A. DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. Nature communications. 2022 Mar 30;13(1):1678.

Yixin Chen et al. Regression Cubes with Lossless Compression and Aggregation. IEEE Trans Knowl Data Eng 18, 1585–1599 (2006).

Challenges in Real-world Data



Non-continuous outcomes

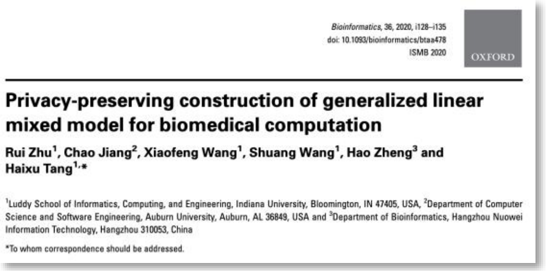
- Binary outcome
- Categorical outcome
- Count outcome
- …

Between-site heterogeneity

We need **Federated Learning Algorithms** for **Generalized Linear Mixed Model (GLMM)**

GLMM Model: $y_{ij} = b_i + \boldsymbol{\beta} x_{\boldsymbol{ij}} + \epsilon_{ij}$, **where** $\epsilon_{ij} \sim N(0, \sigma^2), b_i \sim N(0, \tau^2)$ **is site-specific random effect for k-th site,** β **is fixed effect.**

Existing Works on Federated Learning Algorithms for GLMM



Bioinformatics, 36, 2020, i128–i135
doi: 10.1093/bioinformatics/btaa478
ISMB 2020

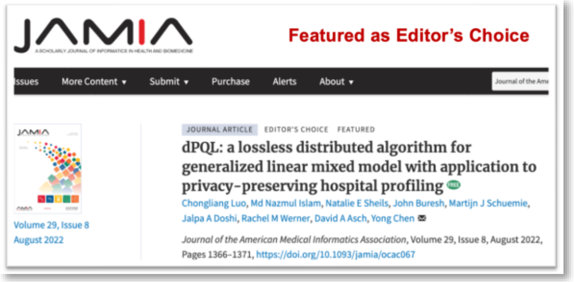
OXFORD

Privacy-preserving construction of generalized linear mixed model for biomedical computation

Rui Zhu[1], Chao Jiang[2], Xiaofeng Wang[1], Shuang Wang[1], Hao Zheng[3] and Haixu Tang[1,*]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47405, USA, [2]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA and [3]Department of Bioinformatics, Hangzhou Nuowei Information Technology, Hangzhou 310053, China

*To whom correspondence should be addressed.



Featured as Editor's Choice

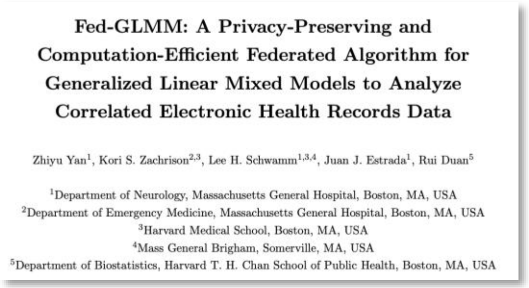
Volume 29, Issue 8
August 2022

JOURNAL ARTICLE EDITOR'S CHOICE FEATURED

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

Chongliang Luo, Md Nazmul Islam, Natalie E Sheils, John Buresh, Martijn J Schuemie, Jalpa A Doshi, Rachel M Werner, David A Asch, Yong Chen

Journal of the American Medical Informatics Association, Volume 29, Issue 8, August 2022, Pages 1366–1371, https://doi.org/10.1093/jamia/ocac067



Fed-GLMM: A Privacy-Preserving and Computation-Efficient Federated Algorithm for Generalized Linear Mixed Models to Analyze Correlated Electronic Health Records Data

Zhiyu Yan[1], Kori S. Zachrison[2,3], Lee H. Schwamm[1,3,4], Juan J. Estrada[1], Rui Duan[5]

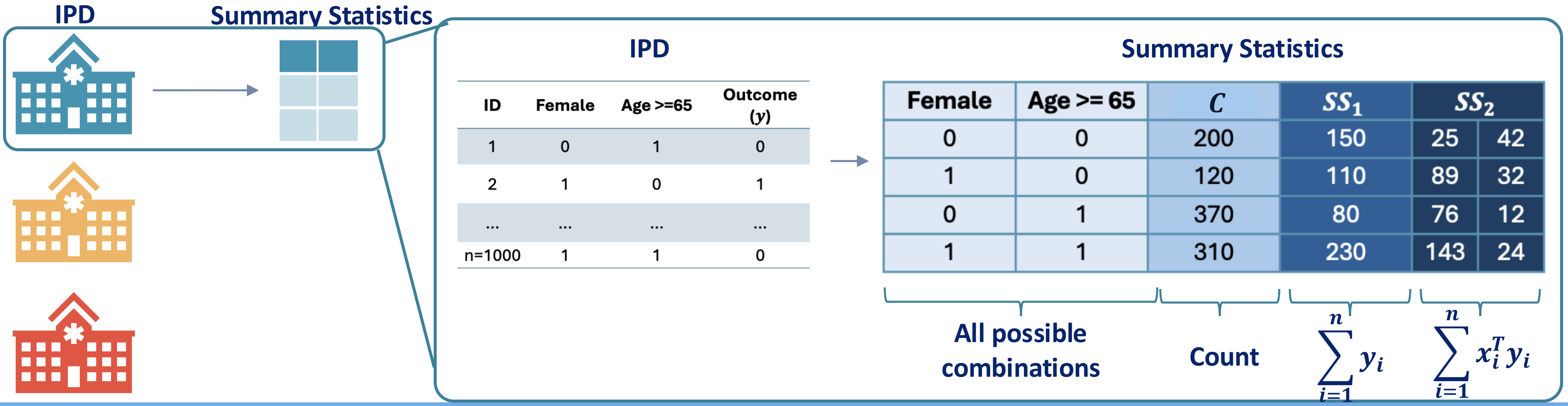
[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[2]Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, USA
[3]Harvard Medical School, Boston, MA, USA
[4]Mass General Brigham, Somerville, MA, USA
[5]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

	Zhu et al, 2020, *Bioinformatics*	Luo et al, 2022, *JAMIA*	Yan et al, 2022, *arxiv*	COLA-GLMM
Lossless	✗	✓	✗	✓
One-shot	✗	✗	i	✓
Communication Round	Iterative (500~1000 rounds)	< 5 rounds	1 or 2 rounds* (Depends on initialization)	1 round

Proposed Method – COLA-GLMM

Collaborative One-shot Lossless Algorithm for Generalized Linear Mixed Model

- Suppose that a **common** set of covariates are available at all collaborating sites.
- The covariates have been **standardized into categorical variables**.
- Pipeline:



IPD

ID	Female	Age >=65	Outcome (y)
1	0	1	0
2	1	0	1
...	...	...	...
n=1000	1	1	0

Summary Statistics

Female	Age >= 65	C	SS_1	SS_2	
0	0	200	150	25	42
1	0	120	110	89	32
0	1	370	80	76	12
1	1	310	230	143	24

All possible combinations

Count

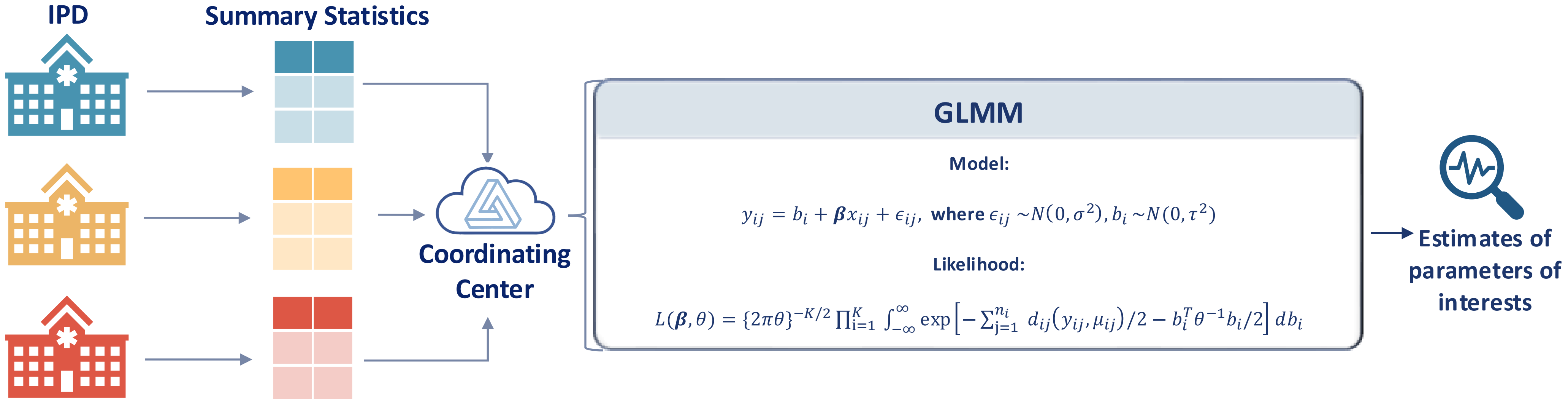
$\sum_{i=1}^{n} y_i$

$\sum_{i=1}^{n} x_i^T y_i$

Proposed Method – COLA-GLMM

Collaborative **O**ne-shot **L**ossless **A**lgorithm for **G**eneralized **L**inear **M**ixed **M**odel

- Suppose that a **common** set of covariates are available at all collaborating sites.
- The covariates have been **standardized into categorical variables**.
- Pipeline:



Model:

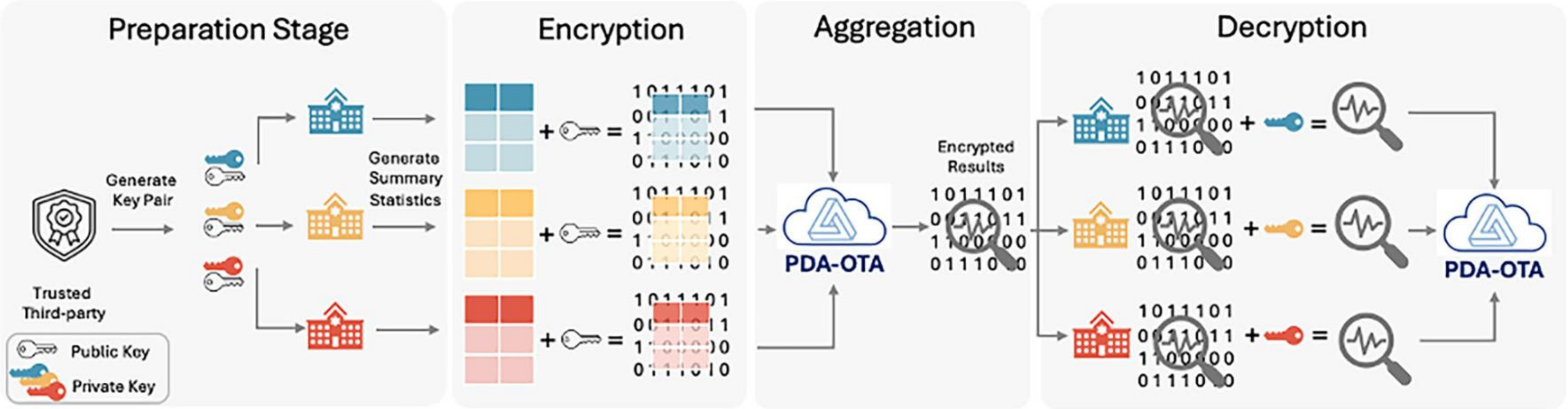
$$y_{ij} = b_i + \boldsymbol{\beta} x_{ij} + \epsilon_{ij},\ \textbf{where}\ \epsilon_{ij} \sim N(0, \sigma^2), b_i \sim N(0, \tau^2)$$

Likelihood:

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

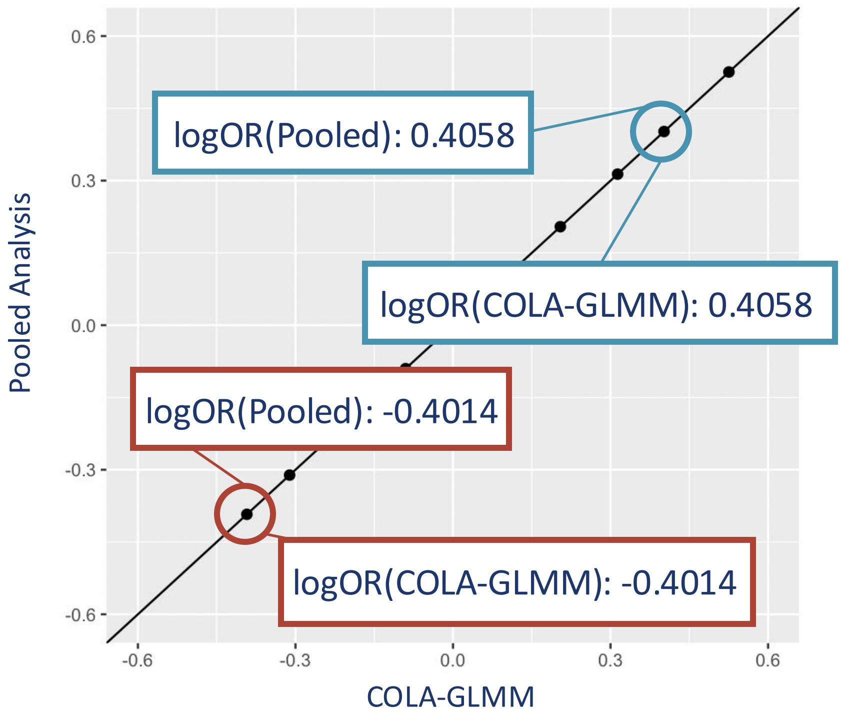
Homomorphic Encryption Enhanced COLA-GLMM

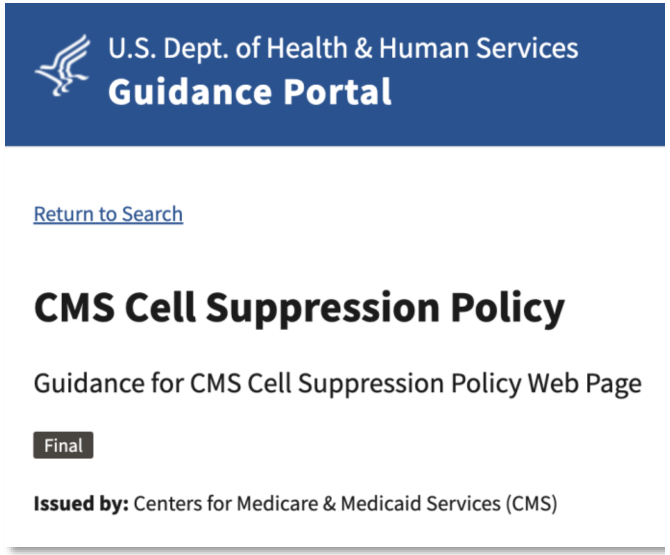
- Under semi-trusted environment:
  - while both data contributors and the coordinating center adhere to the protocol without engaging in malicious actions, the coordinating center may still attempt to derive insights from passively obtained data, indicating a curiosity in extracting information.



Simulation Study – Compare Pooled Analysis and COLA-GLMM

No cell suppression




U.S. Dept. of Health & Human Services
Guidance Portal

Return to Search

CMS Cell Suppression Policy

Guidance for CMS Cell Suppression Policy Web Page

Final

Issued by: Centers for Medicare & Medicaid Services (CMS)

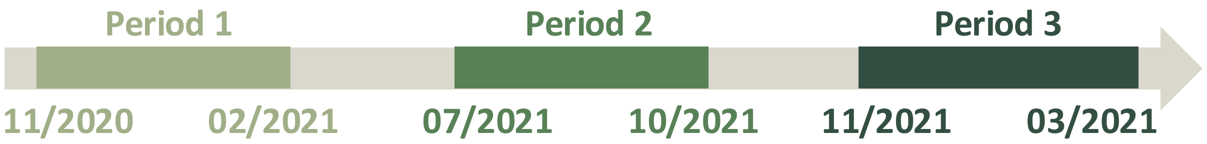
Raw value	Report
0	0
1-10	<11
>=11	as it

Real-world Case Study

- **Scientific Question:**

Identify COVID-19 mortality **risk factors**
over **three time periods** among hospitalized patients

- **Study Period:**



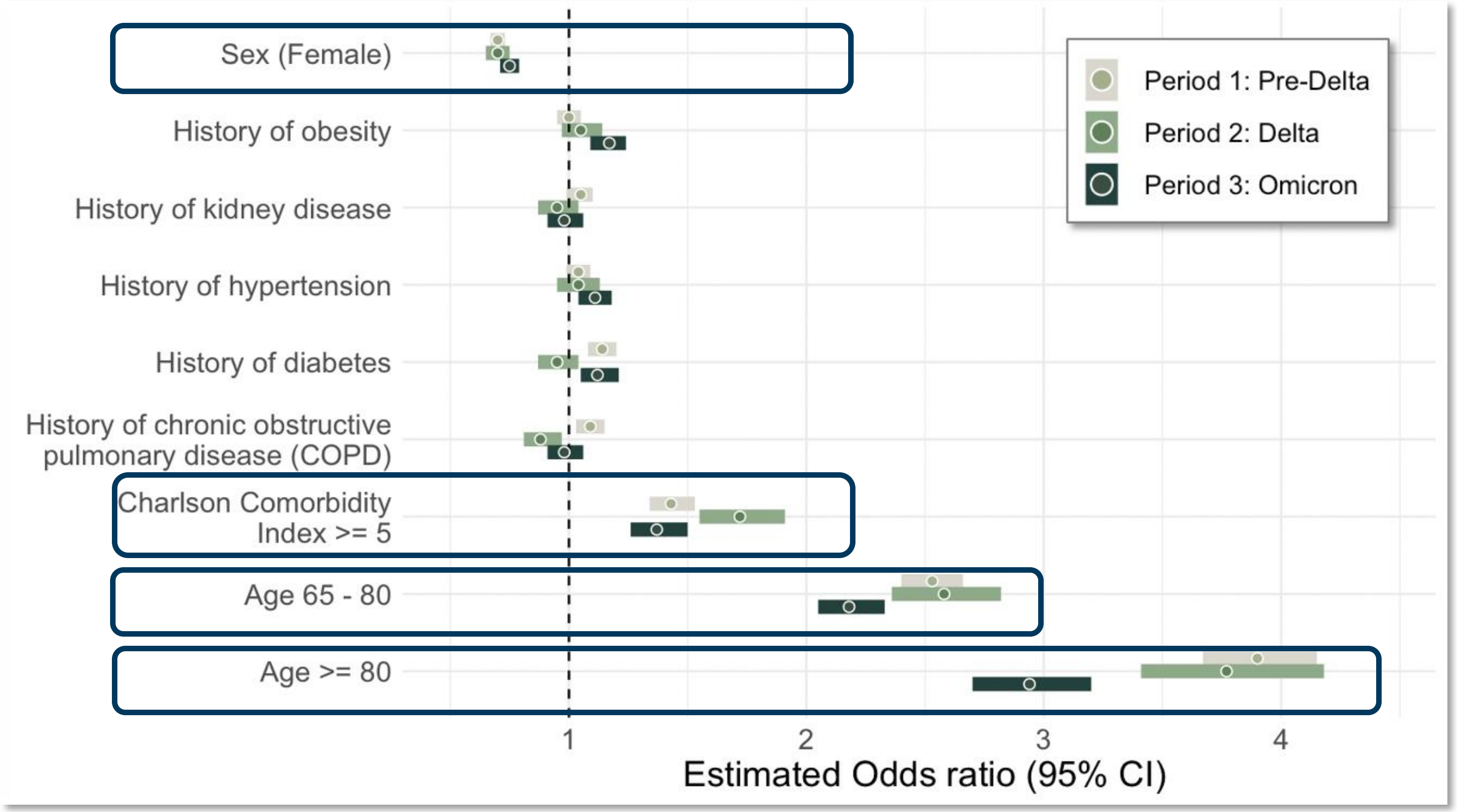
- **Databases (3 countries):**

- Optum® de-identified Electronic Health Record Dataset (Optum EHR);
- Optum's Clinformatics® Data Mart (CDM or Clinformatics®);
- IQVIA Hospital CDM;
- University of Florida Health;
- Department of Veterans Affairs;
- Integrated Primary Care Information (IPCI), The Netherlands;
- Columbia University Irving Medical Center (CUIMC);
- Parc Salut Mar Barcelona (PSMAR), Spain.

Inclusion criteria:

- Patients aged 18 years and older
- Had an inpatient visit with either a diagnosis of COVID-19 or a positive test for COVID-19 between 21 days prior to the inpatient visit and the end of the inpatient visit

Real-world Case Study Results



- **Sex (female):**
  - Reference group: Male
  - Female patients consistently exhibit a lower risk of mortality compared to males across all periods
- **Charlson Comorbidity Index (CCI):**
  - Reference group: CCI < 5
  - Higher CCI scores are statistically associated with an increased risk of mortality.
- **Age:**
  - Reference group: Age < 65
  - Higher age indicates significantly increased risk of mortality

Summary – COLA-GLMM

Collaborative One-shot Lossless Algorithm for Generalized Linear Mixed Model

- **Lossless One-Shot**
- **Summary Statistics Only**
- **Heterogeneity-Aware**
- **Scalable, Applicable, and Implementation-Ready in OHDSI Network**

PDA R Package: 13300+ downloads since 2020

PDA Github Page: https://github.com/Penncil/pda

PDA website: https://pdamethods.org/

PDA-OTA: https://pda-ota.pdamethods.org/

Penn security office certified

Penn Medicine

PennCIL
A Computing · Inference · Learning lab at University of Pennsylvania

CHASE
Center for Health AI & Synthesis of Evidence

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Acknowledgements

- Yong Chen, University of Pennsylvania
- David A. Asch, University of Pennsylvania
- Jenna Reps, Janssen Research and Development
- Chongliang Luo, Washington University in St. Louis
- Yiwen Lu, University of Pennsylvania
- Milou T. Brand, Real World Solutions, IQVIA
- Scott L. DuVall, VA Informatics and Computing Infrastructure
- Thomas Falconer, Columbia University
- Juan Manuel Ramirez-Anguita, Hospital del Mar Research Institute (HMRIB)
- Miguel A. Mayer, Hospital del Mar Research Institute (HMRIB)
- Michael E Matheny, VA Informatics and Computing Infrastructure
- Alex Mayer Fuentes, Parc Taulí Hospital Universitari
- Xing He, University of Florida
- Bhavnisha K Patel, VA Informatics and Computing Infrastructure
- Katherine R Simon, VA Informatics and Computing Infrastructure
- Marc A. Suchard, University of California, Los Angeles
- Guojun Tang, University of Calgary
- Benjamin Viernes, VA Informatics and Computing Infrastructure
- Fei Wang, Weill Cornell Medicine
- Ross D. Williams, Erasmus University Medical Center
- Mui van Zandt, Real World Solutions, IQVIA
- Jiang Bian, University of Florida
- Jiayu Zhou, Michigan State University

Correspondence to:

- **Jessie Tong**, jtong20@jhu.edu
- **Yong Chen**, ychen123@upenn.edu