Distributed Analysis versus Meta-Analysis for Electronic Health Record Data

Andrew Chen
Advised by Jing Huang, Ph.D.

January 15, 2019

Motivation

- Clinical data stored across multiple databases and health systems
- Unable to transfer to central data repository
- Prevents many standard statistical methods

Outline

- Distributed analysis
- Fixed-effect meta-analysis
- Simulation results
- Conclusions
- Future directions

Distributed Analysis

- K sites each with n_k patients, $k = 1, \ldots, K$, with total patients $n = \sum_{k=1}^{K} n_k$
- Incorporates information distributed across multiple sites without pooling data into a single location
- Recent method developed for distributed logistic regression called Grid Binary Logistic Regression (GLORE) (Wu et al. 2012)

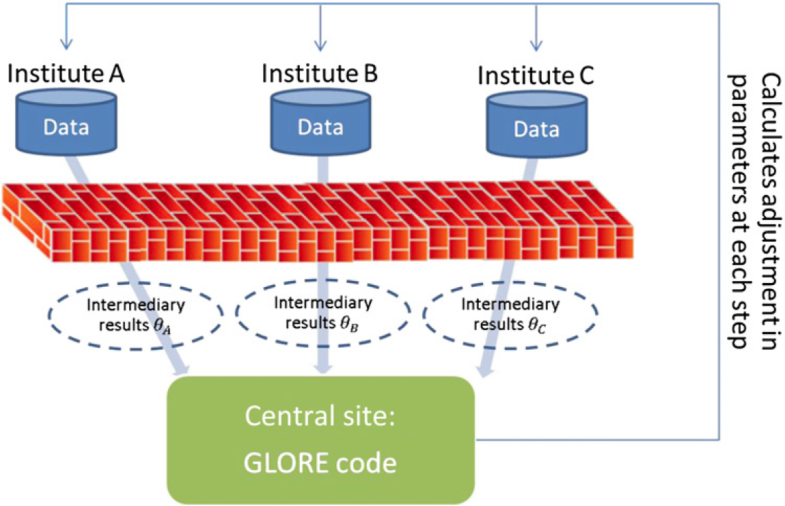
GLORE

- GLORE estimates β via Newton's method

$$\beta^{(j+1)} = \beta^{(j)} - \left[\frac{\partial^2 l(\beta^{(j)})}{\partial \beta^{(j)} \partial \beta^{(j)T}} \right]^{-1} \frac{\partial l(\beta^{(j)})}{\partial \beta^{(j)}}$$

- Both Hessian and gradient can be rewritten as sums of terms calculated from one site each

GLORE


Institute A
Institute B
Institute C
Data
Data
Data
Intermediary results θ_A
Intermediary results θ_B
Intermediary results θ_C
Central site:
GLORE code
Calculates adjustment in parameters at each step

Fixed-Effect Meta-Analysis

- GLORE requires communication between sites for every iteration
  - Original paper found up to six iterations required to reach desired precision
- Meta-analysis provides alternative way to estimate β
- Obtains $\hat{\beta}_k$ and variance-covariance matrix $\hat{V}_k$ for each site
- Then uses inverse-variance weighted estimator for β

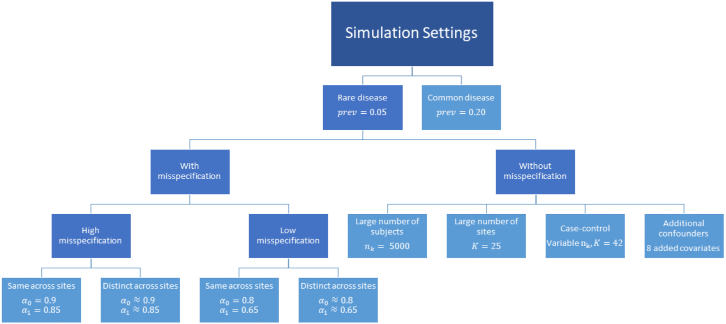
$$\hat{\beta} = \left(\sum_{k=1}^{K} \hat{V}_k^{-1} \right)^{-1} \sum_{k=1}^{K} \hat{V}_k^{-1} \hat{\beta}_k$$

- For fixed K and $n \to \infty$, meta-analysis estimator converges to same limiting distribution as mega-analysis estimator (Lin and Zeng 2010)

Comparison via Simulations

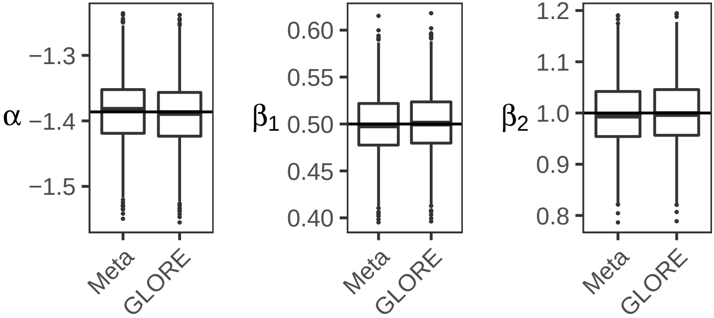
- Default $K = 5$, $n_k = 1000$
- 1000 simulations
- Logistic regression model: $logit(\Pr(Y_i = 1)) = \alpha + \beta X_i$, $i = 1, \ldots, n$
- α depends on case prevalence
- $\beta = (0.5, 1)$
- $X_1 \sim \mathrm{N}(0, 1)$
- $X_2 \sim \mathrm{Bernoulli}(0.5)$
- Newton's method performed with precision of 10^{-6} and starting values of 0

Comparison via Simulations



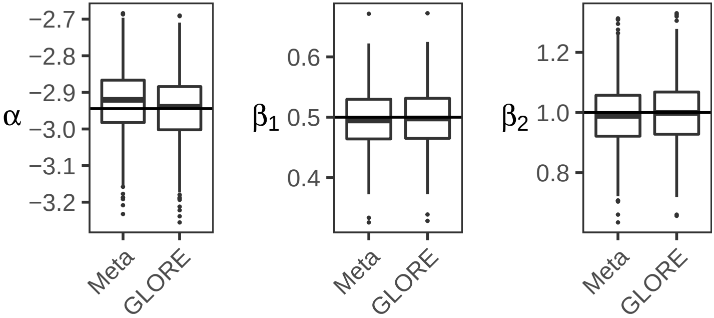
Simulation Results: Common Disease

- Simulated via case prevalence of 20%



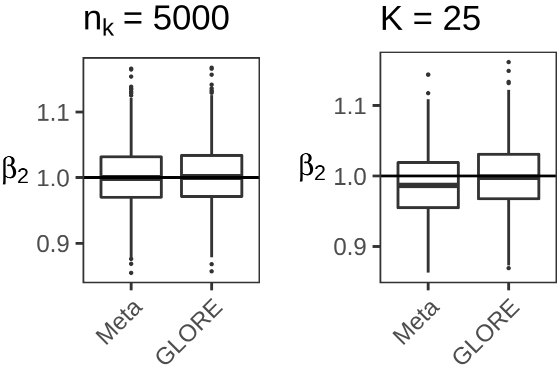
Simulation Results: Rare Disease

- Simulated via case prevalence of 5%



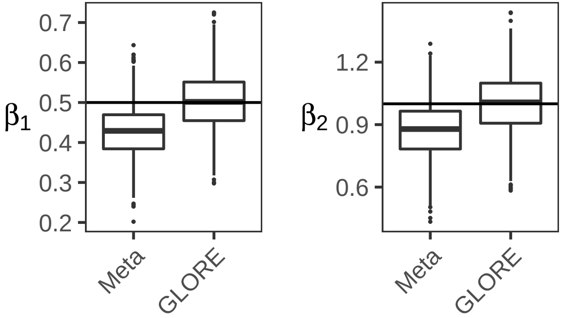
Simulation Results: Large Sample

- Increasing n_k to 5000 brings meta estimates closer to GLORE
- Keeping n_k fixed and increasing K to 25 does not substantially improve meta estimates



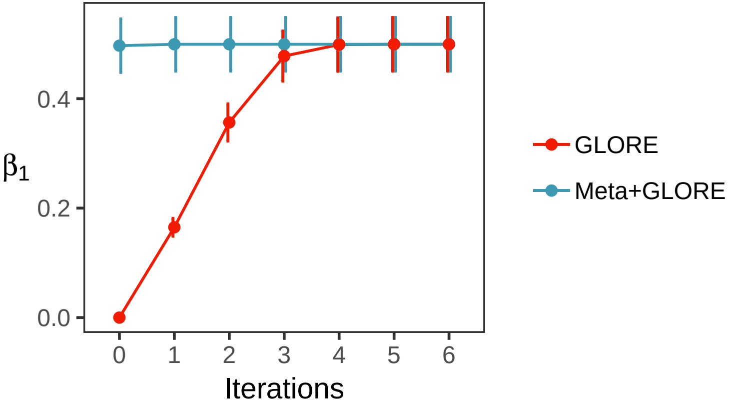
Simulation Results: EHR Settings

- Incorporating misspecified cases or confounders increases bias only slightly
- Case-control setting: 42 sites, two with 1000 subjects and 40 with 75 subjects
  - Sampled cases and 1:1 matched with randomly selected controls



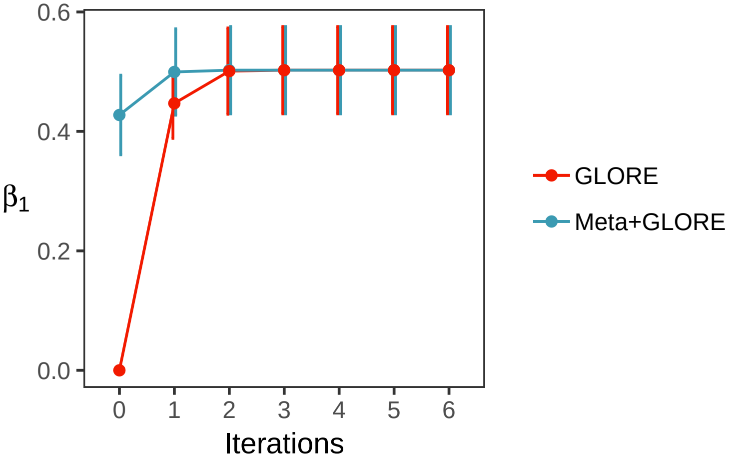
Simulation Results: GLORE Starting Values

- Can use meta-analysis estimates as starting values for GLORE
- In simple rare disease setting, Meta+GLORE requires only 3 iterations to reach desired precision versus 6 for GLORE



Simulation Results: GLORE Starting Values

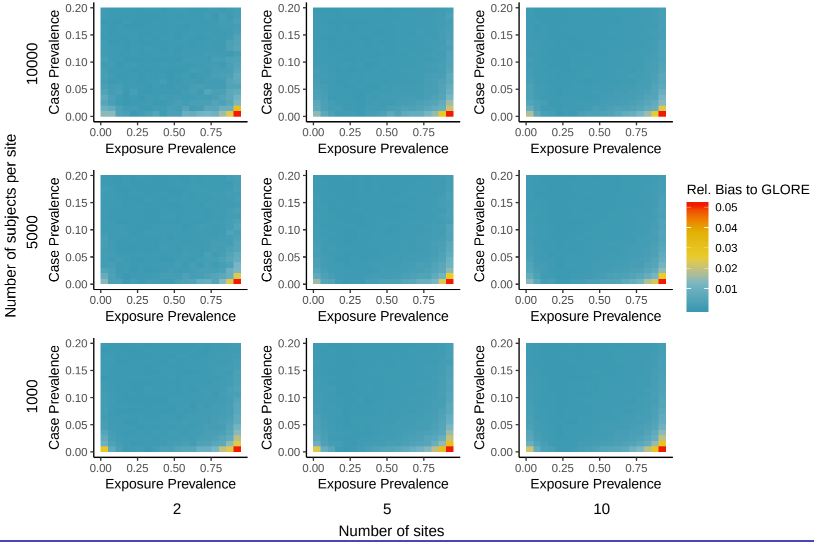
- In case-control setting, Meta+GLORE requires 3 iterations and GLORE requires 4 iterations



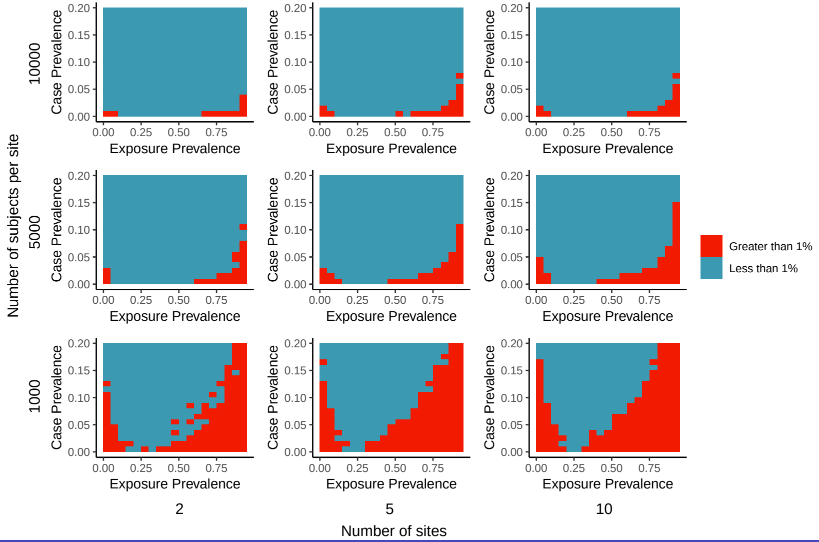
Simulation Results: Many Situations

- Heatmaps to visualize effectiveness of meta-analysis compared to distributed analysis over a grid of simulation parameters
- Case prevalance ranged from 1% to 20%
- Binary exposure prevalence ranged from 5% to 95%
  - Effect size set to 1.0
- 500 simulations for each prevalence pair

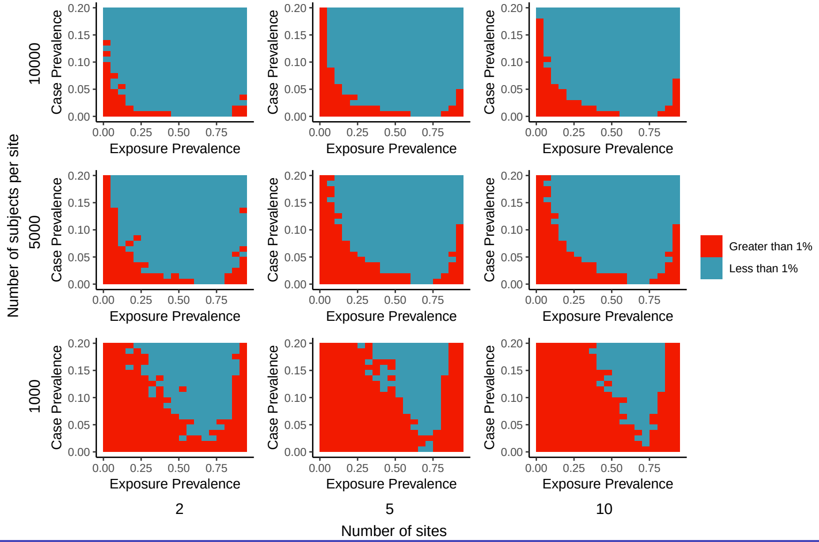
Simulation Results: Many Situations



Simulation Results: Many Situations



Simulation Results: Many Situations, Negative Effect Size



Conclusions

- Meta-analysis shows comparable performance to distributed analysis in many EHR-like settings
- Incorporating meta-analysis estimates can reduce distributed analysis iterations
- Range of settings with acceptable meta-analysis performance depends on number of sites and subjects

Future Directions

- Real data applications
  - Janssen Pharmaceuticals databases
  - Multicenter randomized trial (555 patients over 15 sites) (Ohye et al. 2010)
- Meta-analysis vs. distributed Cox regression (Lu et al. 2015)

Acknowledgments

I would like to thank Jing for her help and mentorship throughout the semester. I would also like to thank Yong and Rui for their contribution of ideas to the project.

References

Lin, D. Y., and D. Zeng. 2010. "On the Relative Efficiency of Using Summary Statistics Versus Individual-Level Data in Meta-Analysis." *Biometrika* 97 (2):321–32.

Lu, Chia-Lun, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, Xiaoqian Jiang, and Lucila Ohno-Machado. 2015. "WebDISCO: A Web Service for Distributed Cox Model Learning Without Patient-Level Data Sharing." *Journal of the American Medical Informatics Association* 22 (6). Oxford University Press:1212–9.

Ohye, Richard G, Lynn A Sleeper, Lynn Mahony, Jane W Newburger, Gail D Pearson, Minmin Lu, Caren S Goldberg, et al. 2010. "Comparison of Shunt Types in the Norwood Procedure for Single-Ventricle Lesions." *New England Journal of Medicine* 362 (21). Mass Medical Soc:1980–92.

Wu, Yuan, Xiaoqian Jiang, Jihoon Kim, and Lucila Ohno-Machado. 2012. "G Rid Binary LO Gistic RE Gression (GLORE): Building Shared Models Without Sharing Data." *Journal of the American Medical Informatics Association* 19 (5):758–64.

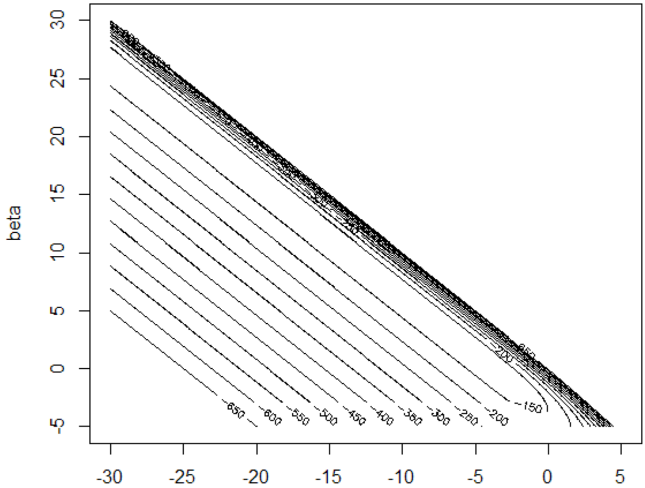
Supplemental: GLORE Derivation

Denote $\pi_i^{(j)} = expit(\alpha^{(j)} + x_i\beta^{(j)})$, and $p_i = \pi_i^{(j)}(1 - \pi_i^{(j)})$. For observations with k variables and for the jth iteration of Newton's method,

$$\left[\frac{\partial^2 l(\alpha^{(j)}, \beta^{(j)})}{\partial(\alpha, \beta)^{(j)}\partial(\alpha, \beta)^{(j)^T}}\right]^{-1} = \begin{bmatrix} \Sigma p_i & \Sigma x_{i1} p_i & \Sigma x_{i2} p_i & \cdots & \Sigma x_{ik} p_i \\ \Sigma x_{i1} p_i & \Sigma x_{i1}^2 p_i & \Sigma x_{i1} x_{i2} p_i & \cdots & \Sigma x_{i1} x_{ik} p_i \\ \Sigma x_{i2} p_i & \Sigma x_{i2} x_{i1} p_i & \Sigma x_{i2}^2 p_i & & \\ \vdots & \vdots & & \ddots & \\ \Sigma x_{ik} p_i & \Sigma x_{ik} x_{i1} p_i & & & \Sigma x_{ik}^2 p_i \end{bmatrix}^{-1}$$

with sums going from $i = 1$ to n, which can be split into sums over individual sites with n_k subjects each

Supplemental: Difficult Situation



Supplemental: Common

Mega (5): -1.39 (-0.00), 0.50 (0.00), 1.00 (-0.00)
Meta (5, 5, 5, 5, 5): -1.38 (0.32), 0.50 (0.43), 1.00 (0.34)
Meta+GLORE 1 (1): -1.39 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 2 (2): -1.39 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 3 (2): -1.39 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 6 (2): -1.39 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 1 (1): -1.16 (16.81), 0.39 (28.89), 0.79 (26.22)
GLORE 2 (2): -1.37 (1.32), 0.49 (1.86), 0.98 (1.68)
GLORE 3 (3): -1.39 (0.01), 0.50 (0.01), 1.00 (0.01)
GLORE 4 (4): -1.39 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 5 (5): -1.39 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 6 (5): -1.39 (0.00), 0.50 (0.00), 1.00 (0.00)

Supplemental: Rare

Mega (6): -2.95 (-0.00), 0.50 (-0.00), 1.00 (-0.00)
Meta (6, 6, 6, 6, 6): -2.93 (0.65), 0.50 (0.51), 0.99 (1.02)
Meta+GLORE 1 (1): -2.95 (0.01), 0.50 (0.01), 1.00 (0.02)
Meta+GLORE 2 (2): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 3 (3): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 6 (3): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 1 (1): -1.78 (39.65), 0.16 (202.68), 0.32 (213.88)
GLORE 2 (2): -2.53 (14.20), 0.36 (40.06), 0.69 (44.70)
GLORE 3 (3): -2.88 (2.34), 0.48 (4.50), 0.94 (5.78)
GLORE 4 (4): -2.94 (0.07), 0.50 (0.11), 1.00 (0.17)
GLORE 5 (5): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 6 (6): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)

Supplemental: Large N

Mega (6): -2.95 (-0.00), 0.50 (0.00), 1.00 (0.00)
Meta (6, 6, 6, 6, 6): -2.94 (0.13), 0.50 (0.10), 1.00 (0.20)
Meta+GLORE 1 (1): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 2 (2): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 3 (2): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 6 (2): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 1 (1): -1.78 (39.65), 0.17 (202.80), 0.32 (213.79)
GLORE 2 (2): -2.53 (14.19), 0.36 (40.02), 0.69 (44.57)
GLORE 3 (3): -2.88 (2.32), 0.48 (4.47), 0.95 (5.70)
GLORE 4 (4): -2.94 (0.07), 0.50 (0.10), 1.00 (0.16)
GLORE 5 (5): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 6 (6): -2.95 (0.00), 0.50 (0.00), 1.00 (0.00)

Supplemental: Large K

Mega (6): -2.94 (0.00), 0.50 (-0.00), 1.00 (-0.00)
Meta (6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6):
-2.92 (0.77), 0.50 (0.62), 0.99 (1.22)
Meta+GLORE 1 (1): -2.94 (0.01), 0.50 (0.01), 1.00 (0.02)
Meta+GLORE 2 (2): -2.94 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 3 (3): -2.94 (0.00), 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 6 (3): -2.94 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 1 (1): -1.78 (39.61), 0.17 (202.68), 0.32 (213.58)
GLORE 2 (2): -2.53 (14.15), 0.36 (39.98), 0.69 (44.49)
GLORE 3 (3): -2.88 (2.31), 0.48 (4.46), 0.95 (5.68)
GLORE 4 (4): -2.94 (0.07), 0.50 (0.10), 1.00 (0.16)
GLORE 5 (5): -2.94 (0.00), 0.50 (0.00), 1.00 (0.00)
GLORE 6 (6): -2.94 (0.00), 0.50 (0.00), 1.00 (0.00)

Supplemental: Case-Control

Mega (4): 0.50 (0.00), 1.00 (0.00)
Meta (4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5): 0.43 (17.54), 0.87 (14.68)
Meta+GLORE 1 (1): 0.50 (0.59), 1.00 (0.42)
Meta+GLORE 2 (2): 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 3 (3): 0.50 (0.00), 1.00 (0.00)
Meta+GLORE 6 (3): 0.50 (0.00), 1.00 (0.00)
GLORE 1 (1): 0.45 (12.42), 0.92 (8.45)
GLORE 2 (2): 0.50 (0.30), 1.00 (0.20)
GLORE 3 (3): 0.50 (0.00), 1.00 (0.00)
GLORE 4 (4): 0.50 (0.00), 1.00 (0.00)
GLORE 5 (4): 0.50 (0.00), 1.00 (0.00)
GLORE 6 (4): 0.50 (0.00), 1.00 (0.00)