Unlocking Efficiency in Real-world Collaborative Studies: A Multi-site International Study with Collaborative One-shot Lossless Algorithm for Generalized Linear Mixed Model

Authors:

Jiayi Tong[a,b,c,*]
Jenna M. Reps[d,e,f]
Chongliang Luo[g]
Yiwen Lu[a,b]
Juan Manuel Ramirez-Anguita[h]
Milou T. Brand[i]
Scott L. DuVall[j,k]
Thomas Falconer[l]
Alex Mayer Fuentes[m]
Xing He[n]
Miguel A. Mayer[h]
Marc A. Suchard[j,o]
Guojun Tang[p]
Ross D. Williams[f]
Fei Wang[q]
Jiang Bian[n]
Jiayu Zhou[r]
David A. Asch[s,t]
Yong Chen[a,b,t,*]

Affiliation of the authors:

a. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
b. The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA
c. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
d. Janssen Research and Development, Titusville, NJ, USA
e. Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA
f. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
g. Division of Public Health Sciences, Department of Surgery, Washington University in St. Louis, St. Louis, MO, USA
h. Hospital del Mar Research Institute (HMRIB), Hospital del Mar, Barcelona, Spain
i. Real World Solutions, IQVIA, Durham, NC, USA
j. VA Informatics and Computing Infrastructure, US Department of Veterans Affairs, Salt Lake City, UT, USA

k. Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA
l. Department of Biomedical Informatics, Columbia University, New York, NY, USA
m. Parc Taulí Hospital Universitari, Sabadell, Spain
n. College of Medicine, Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, FL, USA
o. Department of Biostatistics, University of California, Los Angeles, CA, USA
p. Department of Electrical and Software Engineering, University of Calgary, Canada
q. Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
r. Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA
s. Division of General Internal Medicine, University of Pennsylvania, Philadelphia, PA, USA
t. Leonard Davis Institute of Health Economics, Philadelphia, PA, USA

Background

Largely driven by the widespread adoption of EHRs in U.S. healthcare settings1 of real-world data (RWD) has yielded more robust and generalizable findings, including those that would be impossible to develop with confidence using smaller datasets. These approaches have been facilitated by the development of distributed research networks (DRNs). Centralized DRNs pool patient data in a central database; decentralized DRNs retain data within individual sites, a common structure to enhance data security.

Notable examples of centralized research networks include approximately 60% of the networks within the National Patient-Centered Clinical Research Network (PCORnet)[1,2], which is a network of networks, such as PEDSnet[3], OneFlorida+[4], INSIGHT[5], etc. Beyond the scope of PCORnet, there are additional centralized networks such as IBM Watson Health[6], the All of Us Consortium[7], Flatiron Database[8], Optum[9,10], UK biobank[11], and eMERGE[12]. In recent years, particularly within the context of the COVID-19 pandemic, centralized initiatives like the National COVID Cohort Collaboration (N3C)[13,14] and the RECOVER initiative[15] have been launched. In contrast, about 40% of PCORnet's networks adopt a decentralized research infrastructure, featuring networks such as the PaTH[16] and STAR[17] networks, as well as the Greater Plains Collaborative (GPC)[18]. Other decentralized initiatives aimed at facilitating international studies include The Consortium for Clinical Characterization of COVID-19 by Electronic Health Records (4CE)[19,20], a key player in COVID-19 research, and the Observational Health Data Sciences and Informatics (OHDSI) community[21], a key stakeholder in the European Health Data and Evidence Network (EHDEN)[22].

Despite the conceptual appeal of these DRNs, their use remains logistically challenging because sharing patient-level data across clinical sites typically involves legal agreements, secure file transfers, and repeated back-and-forth communication, each requiring dedicated advocates at each institution. Centralized DRNs have invested extensive effort to streamline these processes. Nonetheless, administrative burden remains large and costly process, adding delay and drag.

To address the challenge in data sharing, the adoption of federated learning algorithms, which enable the conduct of multi-site studies with larger sample sizes without requiring patient-level data sharing. These algorithms also enhance the power and provide more generalizable clinical evidence. This approach is particularly beneficial for decentralized DRNs, especially those involving international collaborators, where sharing patient-level data among hospitals is rarely possible due to the privacy concerns.

Another significant and practical consideration in multi-site analysis is the potential existence of between-site patient heterogeneity. Different sites often attract patients varying considerably in illness severity, comorbidities, social circumstances, and health care needs. This between-site heterogeneity creates confounding bias unless recognized and accounted for. No federated algorithm for GLMM has yet been developed that combines both lossless and one-shot properties. The "lossless" property ensures that results from the federated algorithms align with those obtained in the ideal setting where patient-level data are pooled together for analysis, also known as pooled analysis, which is considered the gold standard. The "one-shot" property refers to achieving results in a single communication round, eliminating the typical back-and-forth data sharing required across collaborating sites and thereby streamlining the process significantly.

To address the methodological gap in existing federated learning algorithms for GLMM that lack the possession of both lossless and one-shot properties, we introduce a novel federated learning algorithm designed to meet the following criteria: (1) requires only summary statistics instead of patient-level data; (2) accounts for between-site heterogeneity at both patient-level and site-level; (3) maintains the lossless property; and (4) achieves results in a single round of communication. Specifically, we propose the Collaborative One-shot Lossless Algorithm for GLMM (COLA-GLMM) algorithm, developed to meet the practical demands of data privacy and efficiency. To assess the performance and applicability of our proposed COLA-GLMM algorithm, we conducted extensive simulation studies and a truly decentralized real-world use-case involving eight data contributors from three countries within the OHDSI network.

Methods

Generalized linear mixed model (GLMM) is an extension of generalized linear model (GLM) with additional random effects. Assume there are K hospitals in total within a network, the k-th site has numbers of patients n_k and the total number of patients within such network is $N = \sum_k n_k$. For subject i at hospital k, we denote y_{ki} the outcome, $\boldsymbol{x}_{ki}$ the p-dimensional covariates with fixed effects $\boldsymbol{\beta}$, and b_k the random effect, $k = 1, \ldots, K, i = 1, \ldots, n_k$. Conditional on the covariates $\boldsymbol{X}_k = \left(\boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kn_k}\right)^T$ and random effects b_k, the outcome $y_k = \left(y_{k1}, \ldots, y_{kn_i}\right)^T$ are assumed to be independent observations with means and variances specified by a generalized linear model as follows:

$$E(y_{ki} \mid b_k) = \mu_{ki} = g(\eta_{ki}) = g\left(\boldsymbol{x}_{ki}^T\boldsymbol{\beta} + b_k\right) \quad (1)$$

$$\mathrm{Var}\ (y_{ki} \mid b_k) = v(\mu_{ki}) \quad (2)$$

where $g(\cdot) = h^{-1}(\cdot)$ is the link function that connects the conditional means μ_{ki} to the linear predictor η_{ki}, and $v(\cdot)$ is the variance function. The random effects b_k are assumed to follow a normal distribution with mean 0 and variance θ (i.e., $b_k \sim N(0, \theta)$).

Consider a *centralized* network where patient-level data from various contributors are aggregated into a central database or data warehouse. To fit GLMM on this multi-site data, the standard procedure for estimating the GLMM parameters ($\boldsymbol{\beta}, \theta$) is through maximizing the integrated quasi-likelihood function, which is written as

$$L(\boldsymbol{\beta}, \theta) = \{2\pi\theta\}^{-K/2} \prod_{k=1}^{K} \int_{-\infty}^{\infty} \exp\left[-\sum_{i=1}^{n_k} d_{ki}(y_{ki}, \mu_{ki})/2 - b_k^T\theta^{-1}b_k/2\right] db_k, \quad (3)$$

where $d_{ki}(y,\mu) = -2\int_{y}^{\mu}(y-u)/v(u)du$. However, the numerical integration techniques required for calculating the likelihood function become exceedingly complex when these are irreducibly high-dimensional integrals. Therefore, the Penalized Quasi-Likelihood (PQL) method was proposed as an approximate approach for estimating parameters in GLMM[23]. The PQL method has proven its suitability for practical applications across various fields[24–26].

Specifically, by applying the Laplace's method for integral approximation, the PQL method leads to iteratively fitting a linear mixed model (LMM). The log-likelihood of LMM with all patient-level data from K sites can be written as:

$$\ell(\boldsymbol{\beta},\theta) = -\frac{1}{2}\sum_{k=1}^{K}\left\{\log|\Sigma_k| + \left(Y_k^* - \boldsymbol{X}_{\boldsymbol{k}}^T\beta\right)^T\Sigma_k^{-1}\left(Y_k^* - \boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{\beta}\right)\right\}, \tag{4}$$

where X_k is the covariate matrix, Y_k^* is the working outcome vector, $|\cdot|$ is the matrix determinant, and $\Sigma_k = \Sigma_k(\theta) = \theta\mathbf{1}_{n_k}\mathbf{1}_{n_k}^T + W_k^{-1}, W_k = \operatorname{diag}\{v(\widehat{\boldsymbol{\mu}}_{\boldsymbol{k}})\}$. In the scenario where we are interested in analyzing the data from a centralized DRN, the above Equation (4) is then fitted on the pooled data to obtain ($\widehat{\boldsymbol{\beta}}$, $\hat{\theta}$). We also refer to such analysis directly using the PQL method on the pooled dataset as the '*pooled analysis*.' This analysis serves as the benchmark (i.e., gold standard) against which we compare the subsequent implementation and empirical evaluation of the proposed federated learning algorithm.

Let $\boldsymbol{x}^{(j)} \in \{0,1\}^p$ denote the j-th unique individual combination $\boldsymbol{x}$, where $j = \{1,\dots,q\}$. All possible combinations can be represented as the $\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\dots,\boldsymbol{x}^{(q)}$. When implementing COLA-GLMM, within a single round of communication, the summary statistics collected from each data contributor consist of a matrix with dimension $q \times (p+2+p)$. The first p columns are all the possible combinations, i.e., $\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\dots,\boldsymbol{x}^{(q)}$. The rest additional columns for the k-th site include:

1. A q-dimensional vector $\boldsymbol{C}_k = \{c_{k1},\dots,c_{kq}\}$, where $c_{kj} = \sum_{i=1}^{n_k} I\left(\boldsymbol{x}_{ki} = \boldsymbol{x}^{(\mathrm{j})}\right)$, counting the number of patients for each combination of $\boldsymbol{x}$.
2. A q-dimensional vector $\boldsymbol{S}_k = \{s_{k1},\dots,s_{kq}\}$, where $s_{kj} = \sum_{i=1}^{n_k} y_{ki}I\left(\boldsymbol{x}_{ki} = \boldsymbol{x}^{(\mathrm{j})}\right)$, representing the sum of observed outcome values y_{ki} for patients who has corresponding combination of $\boldsymbol{x}$.
3. A $q \times p$ dimensional matrix $\boldsymbol{U}_k = \{\boldsymbol{u}_{k1}^T,\dots,\boldsymbol{u}_{kq}^T\}$, where $\boldsymbol{u}_{kj} = \sum_{i=1}^{n_k} \boldsymbol{x}_{\boldsymbol{ki}}^T y_{ki} I\left(\boldsymbol{x}_{ki} = \boldsymbol{x}^{(\mathrm{j})}\right)$, representing the sum of $\boldsymbol{x}_{\boldsymbol{ki}}^T y_{ki}$ for patients who has corresponding combination of $\boldsymbol{x}$.

Once the coordinating center collects the summary statistics matrices from all data contributors, it can construct the likelihood function as shown in Equation (4). This reconstruction enables the estimation of the parameters of interest to obtain ($\widetilde{\boldsymbol{\beta}}$, $\tilde{\theta}$), where $\widetilde{\boldsymbol{\beta}}$, the estimated fixed effect, represent the association between the outcome of interest and the covariates, thereby helping to identify the risk factors. Additionally, we can also estimate the random effect $\tilde{\theta}_{\boldsymbol{k}}$, which capture the heterogeneous site-specific effects, allowing for the quantification of between-site heterogeneity and site-specific predictions.

Simulation studies

To evaluate the performance and accuracy of the proposed COLA-GLMM algorithm, we conducted simulation studies to compare it with the benchmark pooled analysis, which involves aggregating patient-level data from all contributors. The synthetic data were simulated based on summary statistics shared by data contributors from the real-world application. Further details on the data and study cohort are available in the subsequent 'Data Application' section. We utilized a logistic regression model, distributing data across eight sites with random sample sizes ranging from 500 to 50,000. We generated nine binary risk factors with prevalence rates varying from 10% to 60% and modeled the binary response variable by adjusting coefficients from -0.4 to 0.5. The results from the COLA-GLMM algorithm were benchmarked against the gold standard estimates. Additionally, to assess the impact of different cell suppression policies on data sharing, we reported results from the COLA-GLMM algorithm using cell sizes adjusted to 3 for groups ranging from 1 to 5, and to 6 for groups ranging from 1 to 11, following the CMS cell suppression policy[27].

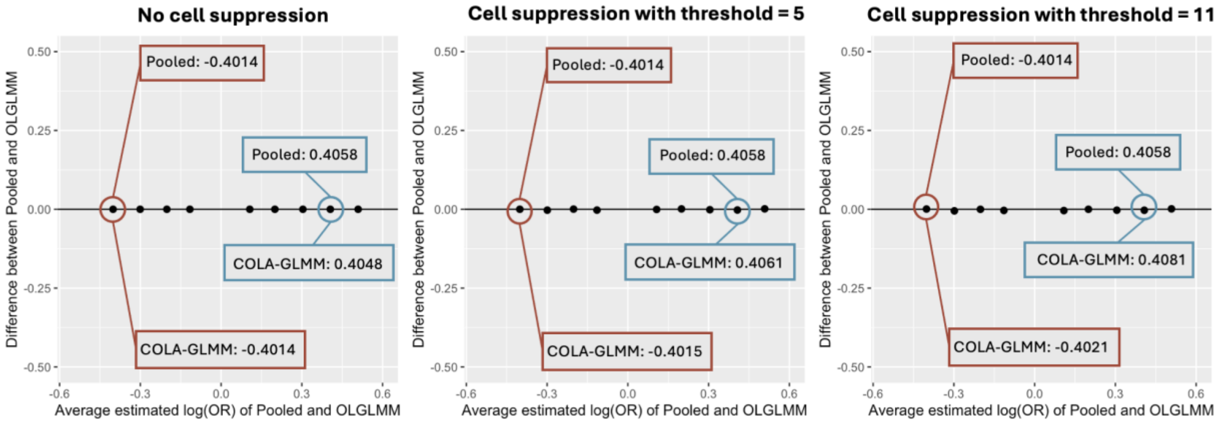

Figure 1. Simulation results visualized via Bland-Alterman plot for comparison between the estimated effect sizes obtained by benchmark (i.e., the pooled analysis) and proposed COLA-GLMM, using different cell suppression

Figure 1 displays a Bland-Altman plot comparing the COLA-GLMM method to the pooled analysis. The y-axis shows the differences in estimated fixed effects on a log(odds ratio) scale, while the x-axis presents the average estimated effect sizes for both methods. Points closer to the horizontal line at zero indicate greater accuracy of the COLA-GLMM relative to the pooled analysis. Two covariates are circled to emphasize their proximity in values, showcasing the precision of the COLA-GLMM method.

Data Application

To demonstrate the applicability of the proposed COLA-GLMM algorithm in a decentralized DRN (i.e., where the patient-level data are not allowed to be shared across sites), we collaborated with eight data contributors from the OHDSI network. We are interested in identifying the risk factors of COVID-19 mortality among hospitalized patients and examining the temporal consistency of these risk factors across three time periods (i.e., pre-Delta, Delta, and Omicron periods or waves) during the COVID-19 pandemic. The data contributors include:

- 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.

Study Cohort & Design

The study cohort included patients aged 18 years and older who 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, with the visit start as the index date. Patients were excluded if they had been observed in the database for fewer than 180 days prior to the index date (i.e., date of hospitalization).The primary clinical outcome was patient death during or up to 7 days after the inpatient visit. The patient-level covariates included age, sex, Charlson Comorbidity Index, history of obesity, Chronic Obstructive Pulmonary Disease (COPD), hypertension, diabetes, and kidney disease. All participating sites standardized their data into the OMOP Common Data Model (CDM).

Implementation

In terms of implementing the framework with all collaborators, a web-based secure platform PDA-OTA (Privacy-preserving Distributed Algorithm Over the Air, https://pda-ota.pdamethods.org/home/) was employed, which enables synchronization of project information and status, allocation of aggregated data (AD), and encryption of hospital information. This platform served as the coordinating center for the collaborating sites to upload and manage the AD. In this study, we adopted a streamlined communication process with only a single round of communication. Each participating site had two main responsibilities: firstly, downloading the control file from the PDA-OTA platform, and secondly, executing an R study package to generate AD using their local patient-level data, followed by uploading the AD to the PDA-OTA platform.

Estimation Results

Figure 2 reports estimated odds ratios (OR) and 95% confidence intervals (CI) for identifying COVID-19 mortality risk factors among hospitalized patients using eight decentralized databases using the COLA-GLMM algorithm. Based on the figure presented, several risk factors have been consistently identified as significant across three study time periods, including:

- **Age:** Being aged 80 and above consistently showed the highest odds ratios among all risk factors across the three periods, indicating a significantly increased risk of mortality. This risk is also notable in the age group of 65-80, though less pronounced than in the older age group.
- **Charlson Comorbidity Index (CCI):** Higher CCI scores are statistically associated with an increased risk of mortality. This association remains consistent across all periods, with a notably stronger correlation during the Delta period.
- **Sex (female):** There is evidence showing that female patients consistently exhibit a lower risk of mortality compared to males across all periods, though the difference is relatively modest.

These findings align with several published studies on risk factor identification[28–31].

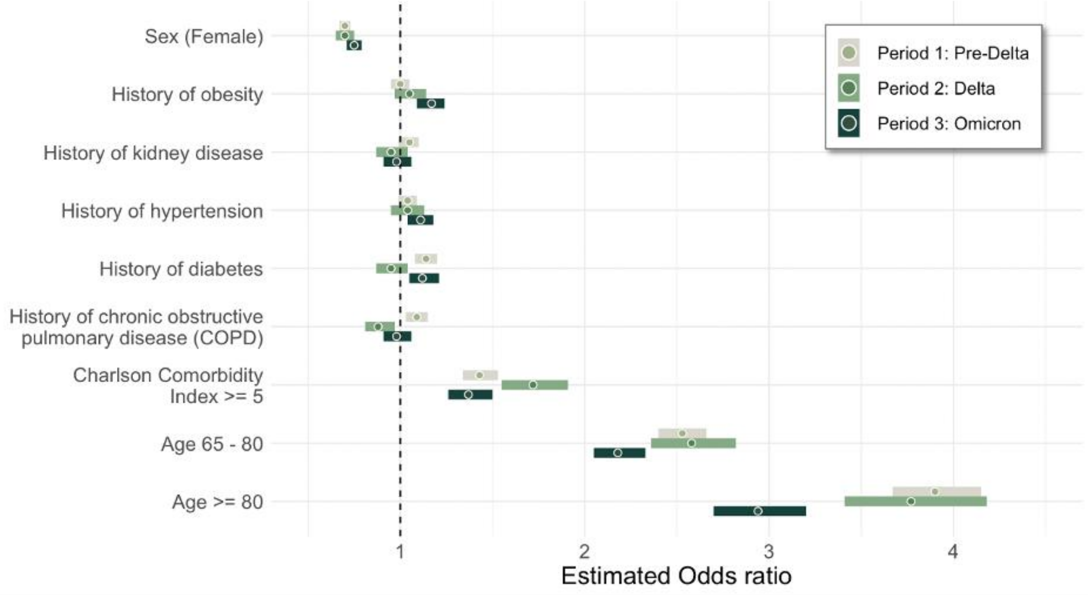

Figure 2: Estimated odds ratios and 95% confidence intervals for identifying COVID-19 mortality risk factors among hospitalized patients using eight decentralized databases using the OGLMM algorithm. *In accordance with the OHDSI cell suppression policy, all cells containing values less than 5 were imputed as 3 when sharing aggregated data across databases.*

Conclusion

We introduced the COLA-GLMM algorithm, a pioneering approach designed for multi-site studies utilizing distributed data without compromising on privacy or accuracy. The algorithm's unique oneshot feature, coupled with its ability to preserve the original data's integrity, marks a significant advancement in the field of federated learning. Throughout our simulation studies and real-world use case, we demonstrated that the COLA-GLMM algorithm not only maintains data confidentiality but also ensures that the aggregated results are equivalent to those obtained from a centralized analysis. This equivalence is critical, as it provides assurance that the privacy-preserving measures do not detract from the analytical value of the data. By requiring minimal rounds of communication and ensuring that data never leaves its original site in an identifiable form, the COLA-GLMM algorithm proves its suitability for sensitive and large-scale studies.

References

1. Collins, F. S., Hudson, K. L., Briggs, J. P. & Lauer, M. S. PCORnet: turning a dream into reality. *J. Am. Med. Inform. Assoc.* **21**, 576–577 (2014).
2. Forrest, C. B. *et al.* PCORnet® 2020: current state, accomplishments, and future directions. *J Clin Epidemiol* **129**, 60–67 (2021).
3. Forrest, C. B. *et al.* PEDSnet: a National Pediatric Learning Health System. *J. Am. Med. Inform. Assoc.* **21**, 602–606 (2014).
4. Shenkman, E. *et al.* OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute With a Community-Based Distributive Medical Education Model. *Acad. Med.* **93**, 451–455 (2018).
5. INSIGHT Clinical Research Network. https://insightcrn.org/.
6. Healthcare | IBM. https://www.ibm.com/industries/healthcare.
7. Investigators, T. A. of U. R. P. The "All of Us" Research Program. *N Engl J Med* **381**, 668 (2019).
8. Flatiron Health | Reimagining the infrastructure of cancer care. https://flatiron.com/.
9. Claims Data | Optum. https://www.optum.com/en/business/life-sciences/real-world-data/claims-data.html.
10. Electronic Health Records (EHR) Data | Optum. https://www.optum.com/en/business/life-sciences/real-world-data/ehr-data.html.
11. UK Biobank - UK Biobank. https://www.ukbiobank.ac.uk/.
12. Electronic Medical Records and Genomics (eMERGE) Network. https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE.
13. Haendel, M. A. *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association* **28**, 427–443 (2021).
14. National COVID Cohort Collaborative | National Center for Advancing Translational Sciences. https://ncats.nih.gov/research/research-activities/n3c.
15. About the Initiative | RECOVER COVID. https://recovercovid.org/.
16. PaTH Network. https://pathnetwork.org/.
17. STAR Clinical Research Network. https://starcrn.org/.
18. GPC – Greater Plains Collaborative. https://gpcnetwork.org/.
19. 4CE. https://covidclinical.net/.
20. Brat, G. A. *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med* **3**, 1–9 (2020).
21. Hripcsak, G. *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* **113**, 7329–7336 (2016).
22. Voss, E. A. *et al.* European Health Data & Evidence Network—learnings from building out a standardized international health data network. *Journal of the American Medical Informatics Association* **31**, 209–219 (2023).
23. Breslow, N. E. & Clayton, D. G. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* **88**, 9–25 (1993).
24. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. S. Viable offspring derived from fetal and adult mammalian cells. *Nature 1997 385:6619* **385**, 810–813 (1997).
25. Sampson, R. J., Raudenbush, S. W. & Earls, F. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science (1979)* **277**, 918–924 (1997).
26. Zhou, W., Nielsen, J., Fritsche, L., … R. D.-N. & 2018, undefined. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *nature.comW Zhou, JB Nielsen, LG Fritsche, R Dey, ME Gabrielsen, BN Wolford, J LeFaiveNature genetics, 2018•nature.com*.
27. CMS Cell Size Suppression Policy. *https://resdac.org/articles/cms-cell-size-suppression-policy*.
28. Noor, F. M. & Islam, M. M. Prevalence and associated risk factors of mortality among COVID-19 patients: a meta-analysis. *J Community Health* **45**, 1270–1282 (2020).
29. Jawad Hashim, M., Alsuwaidi, A. R. & Khan, G. Population risk factors for COVID-19 mortality in 93 countries. *J Epidemiol Glob Health* **10**, 204–208 (2020).
30. Albitar, O., Ballouze, R., Ooi, J. P. & Ghadzi, S. M. S. Risk factors for mortality among COVID-19 patients. *Diabetes Res Clin Pract* **166**, 108293 (2020).
31. Comoglu, S. & Kant, A. Does the Charlson comorbidity index help predict the risk of death in COVID-19 patients? *North Clin Istanb* **9**, 117 (2022).