OHDSI: Applying the Decentralized Generalized Linear Mixed Effects Model (dGEM) for Hospital Profiling of COVID-19 Mortality Data across OHDSI Network

**Version:** 0.1 (feasibility test)

**Authors:**

Yong Chen, Department of Biostatistics, Epidemiology and Informatics (DBEI), the Perelman School of Medicine, University of Pennsylvania

Jiayi Tong, Department of Biostatistics, Epidemiology and Informatics (DBEI), the Perelman School of Medicine, University of Pennsylvania

Jenna Reps, Janssen R&D, Titusville, NJ

**Date:** March 25, 2022

**Acknowledgement:** The analysis is based in part on work from the Observational Health Sciences and Informatics collaborative. OHDSI (<http://ohdsi.org>) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.

Table of contents

[2 List of abbreviations 2](#_Toc98328022)

[3 Abstract 3](#_Toc98328023)

[4 Amendments and Updates 3](#_Toc98328024)

[5 Rationale and Background 3](#_Toc98328025)

[6 Study Objectives 4](#_Toc98328026)

[6.1 Research Questions 4](#_Toc98328027)

[6.2 Objectives 5](#_Toc98328028)

[7 Research methods 6](#_Toc98328029)

[7.1 Study Design 6](#_Toc98328030)

[7.1.1 Overview 6](#_Toc98328031)

[7.2 Study population 6](#_Toc98328032)

[7.3 Outcome 7](#_Toc98328033)

[7.3.1 Mortality 7](#_Toc98328034)

[7.4 Covariates 7](#_Toc98328035)

[8 Data Analysis Plan 8](#_Toc98328036)

[8.1 Creation of Aggregated Data 8](#_Toc98328037)

[8.2 dGEM 8](#_Toc98328038)

[8.2.1 Statistical models 8](#_Toc98328039)

[8.3 Output 8](#_Toc98328040)

[8.4 Data Sources 8](#_Toc98328041)

[8.5 Quality control 8](#_Toc98328042)

[8.6 Strengths and Limitations of the Research Methods 8](#_Toc98328043)

[9 Protection of Human Subjects 9](#_Toc98328044)

[10 Plans for Disseminating and Communicating Study Results 9](#_Toc98328045)

[11 References 9](#_Toc98328046)

[12 Appendix: Study Population Definitions 10](#_Toc98328047)

# List of abbreviations

IPD Individual Patient Data

GLMMs Generalized Linear Mixed Models

dGEM Decentralized Algorithm for Generalized Linear Mixed Effects Model

# Abstract

Hospital profiling, which evaluates how much patient outcomes are influenced by the hospital, allows for a quantitative comparison of healthcare providers' quality of care for certain clinical outcomes (e.g., mortality rate). Given the novelty of COVID-19, the study of hospital profiling with COVID-19 specific data is of great interest. The OHDSI network contains a large number of datasets with COVID-19 data and when combined the COVID-19 data are rather large. However, due to privacy issues, it is not possible to pool the datasets during multi-site collaboration. For example, sensitive individual patient data (IPD) including the patient's identity, diagnoses, and treatments are usually not allowed under privacy regulation to be shared across networks. Additionally, for hospital profiling, hospital-level encryption is also needed to keep the hospital health information safe.

In this study we propose implementing a novel one-shot decentralized algorithm for generalized linear mixed effects models (dGEM). To the best of our knowledge, dGEM is the first real-world hospital profiling solution to account for heterogeneity in multi-site data in a one-shot distributed manner. The proposed algorithm (i.e., dGEM) is based on the generalized linear mixed effect models (GLMM). The dGEM method assumes common fixed-effects of the factors (i.e., patient- and hospital-level factors) and hospital-specific random effects (i.e., random slopes and intercepts) to calculate the directly standardized COVID-19 mortality rates1 for hospital profiling. The proposed method achieves both patient-level privacy protection by only requiring aggregated data; additionally, the hospital-level encryption is accomplished since each hospital can only access their own standardized mortality rate, and the ranking of the hospitals is conducted anonymously using dGEM algorithm.

The aim of this study is to test the performance of the dGEM method for distributed network analyses of hospital profiling on the COVID-19 mortality rate within the OHDSI network. We will implement the dGEM methodology across the COVID-19 datasets within the OHDSI network for the use case of hospital profiling of COVID-19 mortality. We will focus on two time periods for the hospital profiling: i) during the Alpha wave and ii) during the Delta wave. This will demonstrate feasibility and enable us to estimate the effect of various factors of COVID-19 mortality. However, the hospital rankings will be presented anonymously (working ID 1, working ID 2, …) and will not show the names of the OHDSI collaborators’ datasets.

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| 0.1 | March xx 2022 | J Reps | Initial draft |
| 0.2 | March xx 2022 | Various editors | Revised draft |
| 0.3 | April 25 2022 | J Reps | Editing timeline + features |

# Rationale and Background

The OHDSI network contains multiple datasets with COVID-19 data. The majority of the datasets only contain small quantities of COVID-19 data but combined across the network the COVID-19 data are large. Analyzing the pooled OHDSI COVID-19 data may help discover new insights. Unfortunately, it is not possible to pool the OHDSI COVID-19 data due to privacy concerns, as patient-level data cannot generally be shared.

As highly accurate pooling methods are not possible across the OHSDI network due to privacy protection concerns, distributed methods are required instead. The standard meta-analysis is efficient (only requires analyzing a data set once), it protects privacy (patient-level data is not shared), it is suitable for heterogeneous data, but it is not accurate. To the best of our knowledge, there is no *one-shot* distributed algorithm developed for the generalized linear mixed models (GLMM), especially for studying the hospital profiling with multi-site data.

Therefore, we developed a novel one-shot distributed method, dGEM, that can efficiently combine heterogeneous data while preserving the privacy of protected patient-level and hospital-level health information and calculate effect estimates that are equivalent to pooling the data (highly accurate). We would like to implement the dGEM method across the OHDSI network for the hospital profiling problem (see Figure 1) to i) demonstrate the applicability of the proposed dGEM method and ii) investigate the effect estimates of patient- and hospital-level factors of COVID-19 mortality.

Diagram

Description automatically generated

Figure 1. Workflow of the implementation of dGEM algorithm

# Study Objectives

## Research Questions

To implement a distributed method that can obtain the same estimates as pooling the data across the OHDSI network to learn the effect of various factors on the mortality rate for patients infected with COVID-19 and study the hospital profiling for COVID-19 mortality rates during the Alpha wave and Delta wave.

|  |  |
| --- | --- |
| **Patient Covariates** |  |
| Age categories: 18-65, 65-80, and 80 |  |
| Charlson comorbidity categories: 0-1, 2-4, and 5 |  |
| gender |  |
| race |  |
| history of cancer | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20persons%20with%20cancer.json |
| chronic obstructive pulmonary disease (COPD) | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20Persons%20with%20COPD.json |
| heart disease | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20Persons%20with%20heart%20disease.json |
| hypertension | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCOVID%20v1%5D%20Persons%20with%20hypertension.json |
| hyperlipidemia | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20Persons%20with%20hyperlipidemia.json |
| kidney disease | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5Bcovid%20v1%5D%20Persons%20with%20kidney%20disease.json |
| obesity | https://github.com/ohdsi-studies/DistributedLMM/blob/master/inst/cohorts/obesity.json |
| Diabetes | https://github.com/ohdsi-studies/DistributedLMM/blob/master/inst/cohorts/diabetes.json |
| **Site Covariates** |  |
| Number of visits recorded in visit table in 2019 |  |

**Table 1**. List of covariates considered in this study

## Objectives

Primary objective

* To demonstrate a distributed method can be applied to the OHDSI network conduct anonymous hospital profiling

Secondary objectives

* To estimate the pooled effects for patient- and hospital-level factors on COVID-19 mortality during the Alpha wave and Delta wave across heterogeneous datasets within the OHDSI network.

# Research methods

## Study Design

### Overview

This study will be a retrospective, observational cohort study. By ‘retrospective’ we mean the study will use data already collected at the start of the study. By ‘observational’ we mean no intervention will take place in the course of this study. By ‘cohort study’ we mean a study population consisting of patients admitted due to COVID-19.

Suppose there are K sites in total. For each site, denote **X** for the matrix where rows are patients and columns are covariates plus a column of 1s for the intercept, denote **y** is the mortality status vector, and denote **Z** for a vector of hospital-level covariates. Suppose there are p-1 patient-level covariates, q hospital-level covariates, and there are n\_i subjects in the i-th hospital (without the loss of generalizability, assume n\_i = n for simplicity). We note that **X** is a n-by-p matrix, **y** is a n-by-1 vector, **Z** is a q-by-1 vector.

The dGEM method only requires extracting the following aggregated information from the ith site:

1. A p-by-1 vector, , which is the estimated common effects of hospital-level covariates by fitting logistic regression model
2. A p-by-p matrix, the covariance matrix of
3. A scalar, , which is the intercept of logistic regression model with meta estimates, (the global coefficient obtained using from all sites)
4. A scalar, the variance matrix of
5. A q-by-1 vector, **Z**, defined as hospital-level covariates
6. A K-by-1 vector, **P**, defined as counterfactual rates

These values are then used by the dGEM to calculate the directly standardized mortality rates for hospital profiling.

## Study population

Our study population consists of:

Patients who have an inpatient visit with a diagnosis of COVID-19 on or during the visit or a positive test for COVID-19 on or during the visit.

Additional inclusion criteria:

* At least 365 days of observation time prior to the index date
* Aged 18+

We will investigate two different time periods corresponding to the Alpha wave and Delta wave per data set. The dates will be determined by plotting the number of COVID-19 hospitalizations by calendar data per dataset.

The index date is the date of hospitalization.

## Outcome

### Mortality

We will calculate the directly standardized COVID-19 mortality rates of the hospitals

## Covariates

|  |  |
| --- | --- |
| **Patient Covariates** | Link |
| Age categories: 18-65, 65-80, and 80 | NA (standard feature from FeatureExtraction) |
| Charlson comorbidity categories: 0-1, 2-4, and 5 | NA (standard feature from FeatureExtraction) |
| gender | NA (standard feature from FeatureExtraction) |
| race | NA (standard feature from FeatureExtraction) |
| history of cancer | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20persons%20with%20cancer.json |
| history of chronic obstructive pulmonary disease (COPD) | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20Persons%20with%20COPD.json |
| history of heart disease | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20Persons%20with%20heart%20disease.json |
| history of hypertension | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCOVID%20v1%5D%20Persons%20with%20hypertension.json |
| history of hyperlipidemia | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5BCovid%20v1%5D%20Persons%20with%20hyperlipidemia.json |
| history of kidney disease | https://github.com/ohdsi-studies/Covid19PredictionStudies/blob/master/CovidSimpleModels/inst/cohorts/%5Bcovid%20v1%5D%20Persons%20with%20kidney%20disease.json |
| history of obesity | https://github.com/ohdsi-studies/DistributedLMM/blob/master/inst/cohorts/obesity.json |
| History of diabetes | https://github.com/ohdsi-studies/DistributedLMM/blob/master/inst/cohorts/diabetes.json |
| **Site Covariates** |  |
| Number of visits recorded in visit table in 2019 |  |

We will use data prior to hospitalization to construct the predictors using the same definitions as previously used to develop a prognostic model in influenza patients [2].

# Data Analysis Plan

## Feasibility

In this study we aim to perform the analysis during i) the Alpha wave and ii) the Delta wave. However, these waves occurred at different time periods for each country. There will be a preliminary feasibility step where each site creates the COVID-19 hospitalization cohort and generates a plot with the date on the x-axis and number of patients in the cohort on the y-axis. This plot will then be inspected by the data holder to identify the start and end of the Alpha and Delta waves in the specific dataset. These dates will then be input for the main analysis. By default, these dates will correspond to the USA Alpha and Delta wave dates.

The feasibility step will be executed during network site enrollment.

## Creation of Aggregated Data

We will create a study package that extracts the patient-level data locally and then calculates the aggregate data components. These will be saved as json files for the collaborator to inspect and automatically extracted using the PDA-OTA (Privacy-preserving Distributed Algorithm Over The Air) platform for sharing. We will then provide instructions for sharing the final results on the PDA-OTA platform.

## dGEM

### Statistical models

Once the aggregate data are collected across the OHDSI network we will implement the dGEM algorithm.

The dGEM is a decentralized algorithm for the generalized linear mixed-effects model (GLMM). It works by conducting a fixed-effect meta-analysis of the common patient-level effects and a random-effects analysis of the random hospital-level effects. The directly standardized mortality rates, with the idea of counterfactual modeling, are calculated for the hospitals based on the estimated effects. Hospital profiling is then conducted based on the directly standardized mortality rates.

## Output

The output of this study will be the directly standardized mortality rates of the hospitals, anonymous hospital ranking, and the pooled effect estimates for each covariate.

## Data Sources

The analyses will be performed across a network of observational healthcare databases. All databases have been transformed into the OMOP Common Data Model, version 5. The complete specification for OMOP Common Data Model, version 5 is available at: <https://github.com/OHDSI/CommonDataModel>.

## Quality control

We will evaluate the aggregate data by

* Performing tests to ensure each predictor is extracted correctly
* Performing tests to ensure the matrix multiplication is implemented correctly
* Testing the study package on a single database to ensure all 4 components are extracted and saved

The PatientLevelPrediction and FeatureExtraction packages, as well as other OHDSI packages on which these depends, use unit tests for validation.

## Strengths and Limitations of the Research Methods

Strength

* This study will enable pooled effects to be estimated across the OHDSI network
* It only requires extracting aggregate data one per site
* It is suitable for heterogenous data
* It achieves health information protection and encryption for both patient- and hospital-level.
* It ensures the reliability of hospital profiling

Limitations

* Many datasets in OHDSI lack specific dates for hospital events and it is not possible to discriminate between patients hospitalized due to COVID-19 and those who catch COVID-19 during hospitalization
* Many OHDSI datasets have incomplete death records and only contain inpatient death, however as our outcome is death during hospitalization, this is less of an issue.
* Race is not well captured across the OHDSI datasets
* The sensitivity/PPV of the predictor phenotypes may differ across the datasets
* As this is a demonstration we only include predictors that have been previously identified

# Protection of Human Subjects

The study is using only de-identified data. Confidentiality of patient records will be maintained at all times. All study reports will contain aggregate data only and will not identify individual patients or physicians.

# Plans for Disseminating and Communicating Study Results

The study results will be posted on the OHDSI website after completion of the study. At least one paper describing the study and its results will be written and submitted for publication to a peer-reviewed scientific journal.

# References

1 Asch DA, Sheils NE, Islam MN, Chen Y, Werner RM, Buresh J, *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**:471–8.

# Appendix: Study Population Definitions

[add]