OHDSI: Applying the Distributed Linear Mixed Model to integrate heterogeneous COVID-19 hospitalization data across the 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

Jenna Reps, Janssen R&D, Titusville, NJ

Martijn Schuemie, Janssen R&D, Titusville, NJ

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

….....

**Date:** August 12, 2020

**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](#_Toc48143106)

[3 Abstract 3](#_Toc48143107)

[4 Amendments and Updates 3](#_Toc48143108)

[5 Rationale and Background 3](#_Toc48143109)

[6 Study Objectives 4](#_Toc48143110)

[6.1 Research Questions 4](#_Toc48143111)

[6.2 Objectives 4](#_Toc48143112)

[7 Research methods 5](#_Toc48143113)

[7.1 Study Design 5](#_Toc48143114)

[7.1.1 Overview 5](#_Toc48143115)

[7.2 Study population 5](#_Toc48143116)

[7.3 Outcome 5](#_Toc48143117)

[7.3.1 Length of hospitalization 5](#_Toc48143118)

[7.4 Covariates 5](#_Toc48143119)

[8 Data Analysis Plan 6](#_Toc48143120)

[8.1 Creation of AD 6](#_Toc48143121)

[8.2 DDLMs 6](#_Toc48143122)

[8.2.1 Statistical models 6](#_Toc48143123)

[8.3 Output 6](#_Toc48143124)

[8.4 Data Sources 6](#_Toc48143125)

[8.5 Quality control 6](#_Toc48143126)

[8.6 Strengths and Limitations of the Research Methods 7](#_Toc48143127)

[9 Protection of Human Subjects 7](#_Toc48143128)

[10 Plans for Disseminating and Communicating Study Results 7](#_Toc48143129)

[11 References 7](#_Toc48143130)

[12 Appendix: Concept Set Definitions 9](#_Toc48143131)

# List of abbreviations

IPD individual patient data

LMMs Linear mixed models

DLMMs distributed linear mixed models

# Abstract

Due to the novelty of COVID-19 many data sets only contain small quantities of COVID-19 specific data. 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.

In this study we propose implementing a novel algorithm, distributed linear mixed models (DLMMs), that is able to learn the coefficients across heterogeneous data and only requires extracting aggregated data from each data site in the network once. Linear mixed models (LMMs) are commonly used in many areas including epidemiology for analyzing multi-site data with heterogeneity. The model assumes site-specific random effects of the covariates (and intercept) on a continuous outcome. To the best of our knowledge, there is no existing approach for fitting LMMs in a distributed manner.

The aim of this study is to test the performance of DLMMs methods for distributed network analyses of prediction models for severe COVID-19 infection in the OHDSI network. We will implement the DLMMs methodology across the COVID-19 datasets within the OHDSI network to estimate the effect of various predictors on length of hospitalization stay (a proxy for severity of COVID-19 infection) that were identified as predictors of severity during the OHDSI COVID-19 study-a-thon.

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| 0. 1 | August 12 2020 | J Reps | Initial draft |
| 0.2 | September 23 2020 | Various editors | Revised draft |

# 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. A previously distributed lossless methods [1] has been shown to be privacy protecting, highly accurate and efficient but it assumes homogeneity of combined data.

A novel distributed method, the DLMMs, has been developed that can efficiently combine heterogeneous data while preserving the privacy of protected health information and calculate effect estimates that are equivalent to pooling the data (highly accurate). In our previous study, DLMMs was implemented using US data and focusing on the outcome of length of stay for patients hospitalized due to COVID-19. The DLMMs was shown the find equivalent effect estimates when using distributed hospital data compared to applying standard methods using pooled data. We would now like to implement the DLMMs across the OHDSI network for the same prediction problem “in patients hospitalized due to COVID-19, predict length of hospital stay” to i) demonstrate it is feasible and ii) investigate the effect estimates of previously identified predictors of severe COVID-19.

# 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 predictors on the duration of hospitalization for patients with COVID-19.

|  |  |
| --- | --- |
| Predictor |  |
| 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 |

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

## Objectives

Primary objective

* To demonstrate a distributed method can be applied to the OHDSI network to get the pooled effects across heterogeneous datasets

Secondary objectives

* To estimate the pooled effects for each predictor on the duration of hospitalization due to COVID-19 across 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 hospitalized due to COVID-19 will be followed from index date (start of hospitalization due to COVID-19) until end of hospitalization.

For each site, denote X for the matrix where rows are patients and columns are predictors, and denote **y** is the length-of-stay vector. Suppose there are p predictors, and there are n\_i subjects. We note that X is a n-by-p matrix, and y is a n-by-1 vector.

The DLMM method only requires extracting the following aggregated data (AD) from the ith site:

1. Sample size n\_i for the i-th site, i.e., the number of patients.
2. A p-by-p matrix, defined as X^T \* X, i.e., transpose of X multiplied by X.
3. A p-by-1 vector, defined as X^T \* y, i.e., transpose of X multiplied by y.
4. A scaler, defined as y^T \* y

These values are then used by the DLMMs to calculate the pooled effect of each predictor on the length of hospitalization.

## 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+

The index date is the date of hospitalization.

## Outcome

### Length of hospitalization

We will predict the duration of the visit containing the COVID-19 diagnosis or positive test,

## Covariates

|  |  |
| --- | --- |
| Predictor | 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 |

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

## Creation of AD

We will create a study package that extracts the patient-level data locally and then calculates the four aggregate data components. These will be saved as CVS files for the collaborator to inspect and automatically extracted into a compressed directory for sharing. We will then provide instructions for sending the results via FSTP.

## DLMMs

### Statistical models

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

The DLMM is a distributed linear mixed-effects model. It works by reconstructing the profile log-likelihood using the aggregate data only. A linear model is used as it can account for the heterogenous effects across data sites.

## Output

The output of this study will be the pooled effect estimates for each predictor.

## 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 influenza patients to ensure all 4 components are extracted and saved

The FeatureExtraction package itself, as well as other OHDSI packages on which FeatureExtraction 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 is privacy protecting – no patient-level data will be shared

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
* Hospitalization length of stay may not be well captured across the OHDSI datasets
* 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. Chen, Y., Dong, G., Han, J., Pei, J., Wah, B.W. and Wang, J., 2006. Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, *18*(12), pp.1585-1599.

2. Williams, R.D., Markus, A.F., Yang, C., Salles, T.D., Duvall, S.L., Falconer, T., Jonnagaddala, J., Kim, C., Rho, Y., Williams, A. and Alberga, A., 2020. Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. *medRxiv*.

# Appendix: Study Population Definitions

[add]