Name: Liz Limpoco

She

Country: Belgium

Affiliation: Hasselt University

Function: PhD Candidate (Statistics)

Main expertise (1-2 lines):

Federated data analysis, statistical modeling

Estimating statistical models when individual patient data are not accessible

Supervisors:

Prof. dr. Christel Faes

Prof. dr. Niel Hens

DSI DATA SCIENCE INSTITUTE

UHASSELT

WWW.UHASSELT.BE/DSI

The Situation: COVID-19 Pandemic

Fragmented Data

Vaccination records are siloed across diverse national and regional health systems

Urgent Decision-making

Need for real-time large-scale evidence to make informed decisions

Data privacy

Privacy laws limit sharing of Individual Patient Data (IPD)

Our solution:
Federated data analysis

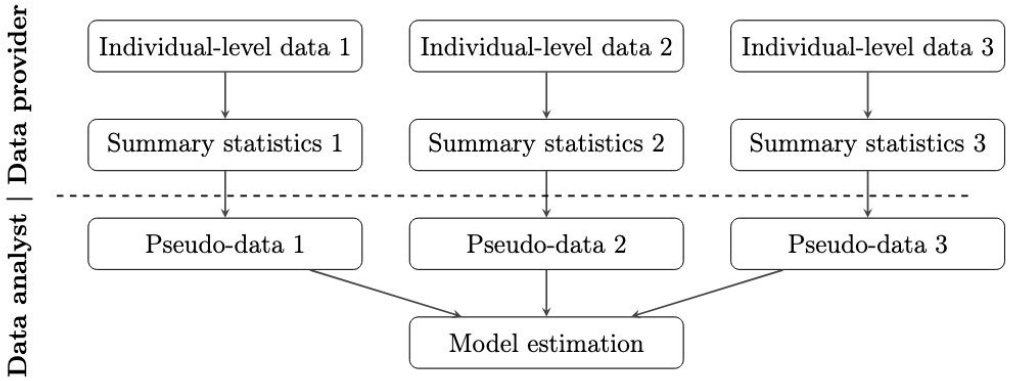


Figure 1: Proposed framework for a setup with three data providers. Each data provider prepares a summarized version of their data. The data analyst generates pseudo-data whose summary statistics match those supplied by the data providers. A global model can then be estimated from the pooled pseudo-data.

summary statistics required:

sample moments (univariate and multivariate) up to 4th order

- **IPD remains federated, not centralized**

  Data providers keep Individual Patient Data (IPD) locally

- **Communication efficient**

  Only summaries are shared once

- **Equivalent to pooled version**

  Statistical models estimated from federated data are equivalent in performance to models estimated from centralizing or pooling all data

Illustrative Example: SPARCS 2022

- **SPARCS inpatient de-identified file**

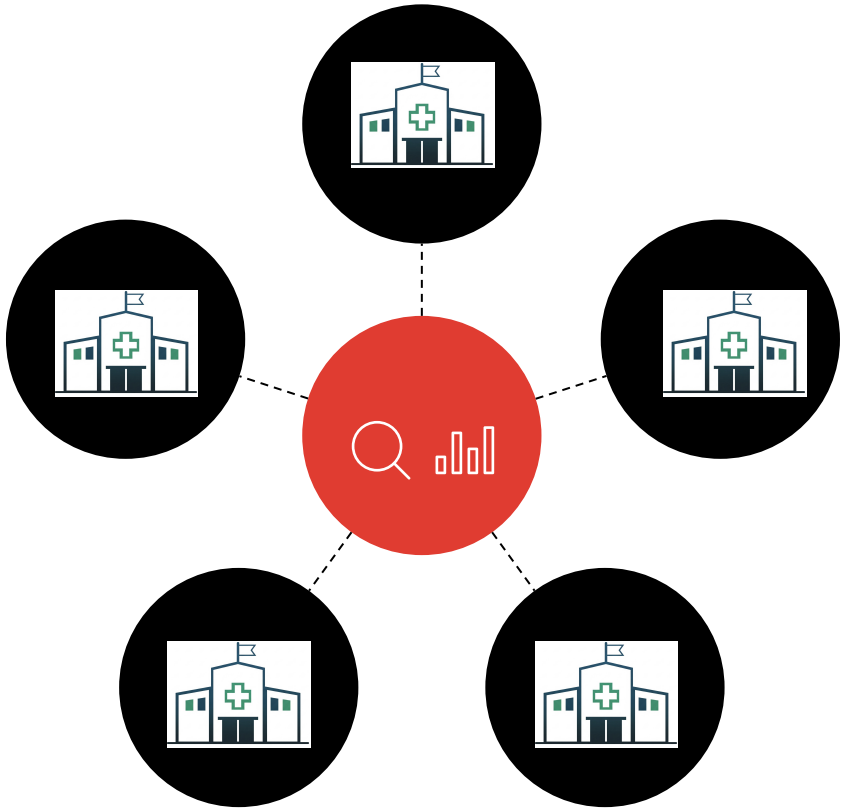
  Statewide Planning and Research Cooperative System (https://health.data.ny.gov) provided by New York State Department of Health

- **Records included in analysis**

  205 hospitals, 2 095 246 patients

- **Variables included in analysis**

  - Sex (M / F)
  - Length of hospital stay (days)
  - COVID-19 test (- / +)
  - Emergency department indicator (N / Y)
  - Total Charges (USD)



Linear [mixed] models

Table 5: Linear mixed model with Total Charges (USD) as response

	pseudo-data		actual data	
	Est(std. err.)	95% CI (Wald)	Est(std. err.)	95% CI (Wald)
(Intercept)	-0.198 (0.031)	(-0.2594, -0.1371)	-0.198 (0.031)	(-0.2594, -0.1371)
Std. Length of Stay (Days)	0.694 (4.7e-04)	(0.6931, 0.6949)	0.694 (4.7e-04)	(0.6931, 0.6949)
COVID19 (Positive)	-0.039 (0.003)	(-0.0443, -0.0327)	-0.039 (0.003)	(-0.0443, -0.0327)
Gender (Male)	0.047 (9.3e-04)	(0.0451, 0.0487)	0.047 (9.3e-04)	(0.0451, 0.0487)
Emergency (Yes)	-0.083 (0.001)	(-0.0853, -0.0812)	-0.083 (0.001)	(-0.0853, -0.0812)
σ_{Int}	0.445		0.445	
AIC	4248016		4248006	
BIC	4248104		4248094	
total number of patients	2,095,246		2,095,246	
number of hospitals	205		205	

$$E[y_{ij}|\mathbf{x}_{ij}] = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + \epsilon_{ij}$$

y_{ij} - Std. Total Charges (USD) per patient j in hospital i
$\mathbf{x}_{ij}$ - vector of predictor data per patient j in hospital i
$\boldsymbol{\beta}$ - vector of fixed effects
u_i - random intercept per hospital i
ϵ_{ij} - normally distributed random error

- **Identical point and interval estimates**
- **Interpretation**
  - + ~ 8 days Length of Stay ⇒ 94 212.30 USD (+)
  - Positive COVID-19 ⇒ 5 294.35 USD (-)
  - Male ⇒ 6 380.37 USD (+)
  - Emergency ⇒ 11 267.47 USD (-)
  - Intraclass Correlation Coefficient (ICC) = 30.78% (*proportion of total variance in y explained by the differences in clusters*)

Binary logistic [mixed] models

Table 6: Logistic mixed model with COVID19 status (Positive or Negative) as response

	pseudo-data		actual data	
	Est(std. err.)	95% CI (Wald)	Est(std. err.)	95% CI (Wald)
(Intercept)	-3.752 (0.071)***	(-3.8913, -3.6125)	-3.752 (0.071)***	(-3.8913, -3.6125)
Gender (Male)	0.163 (0.009)***	(0.1458, 0.1806)	0.163 (0.009)***	(0.1458, 0.1806)
σ_{Int}	0.987		0.987	
AIC	484499.6		484499.5	
BIC	484537.3		484537.2	
total number of patients	2,095,246		2,095,246	
number of hospitals	205		205	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

$$\text{logit}\,(E[y_{ij}|\mathbf{x}_{ij}]) = \mathbf{x}_{ij}^{T}\boldsymbol{\beta} + u_i + \epsilon_{ij}$$

y - COVID-19 test result per patient j in hospital i
$\mathbf{x}_{ij}$ - vector of predictor data per patient j in hospital i
$\boldsymbol{\beta}$ - vector of fixed effects
u_i - random intercept per hospital i
ϵ_{ij} - normally distributed random error

- **Identical point and interval estimates**
  (up to 3rd decimal place)
- **Interpretation**
  - Male ⇒ 17.70% higher odds of being COVID-19 positive than females
  - Intraclass Correlation Coefficient (ICC) = 22.84% (*proportion of total variance in y explained by the differences in clusters*)

Poisson [mixed] models

Table 7: Poisson mixed model with Length of Hospital Stay (Days) as response

	pseudo-data		actual data	
	Est(std. err.)	95% CI (Wald)	Est(std. err.)	95% CI (Wald)
(Intercept)	1.487 (0.034)***	(1.4211, 1.5524)	1.487 (0.033)***	(1.4212, 1.5525)
COVID19 (Positive)	0.158 (0.002)***	(0.1544, 0.1610)	0.158 (0.002)***	(0.1544, 0.1610)
Gender (Male)	0.124 (5.8e-04)***	(0.1229, 0.1252)	0.124 (5.8e-04)***	(0.1229, 0.1252)
Emergency (Yes)	0.318 (6.9e-04)***	(0.3167, 0.3194)	0.318 (6.9e-04)***	(0.3167, 0.3194)
σ_{Int}	0.478		0.478	
(truncated) AIC	(-18965154)		(-18965147) 18521772	
(truncated) BIC	(-18965092)		(-18965084) 18521834	
total number of patients	2,095,246		2,095,246	
number of hospitals	205		205	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

$$\log\left(E[y_{ij}|\mathbf{x}_{ij}]\right) = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + u_i + \epsilon_{ij}$$

y - Length of Stay (days) per patient j in hospital i
$\mathbf{x}_{ij}$ - vector of predictor data per patient j in hospital i
$\boldsymbol{\beta}$ - vector of fixed effects
u_i - random intercept per hospital i
ϵ_{ij} - normally distributed random error

- **Identical point and interval estimates**
  (up to 3rd decimal place)
- **Interpretation**
  - Positive COVID-19 ⇒ 17.12% longer LOS
  - Male ⇒ 13.20% longer LOS
  - Emergency ⇒ 37.44% longer LOS
  - Variance of random intercept: Baseline LOS varies across hospitals

Research questions that can be answered

- Identify significant risk factors
- Quantify relationship between predictors and a response variable
- Provide confidence intervals
- Account for heterogeneity across different data providers
- Perform model selection via AIC
- Generate predictions

Equivalence depends on matching summary statistics

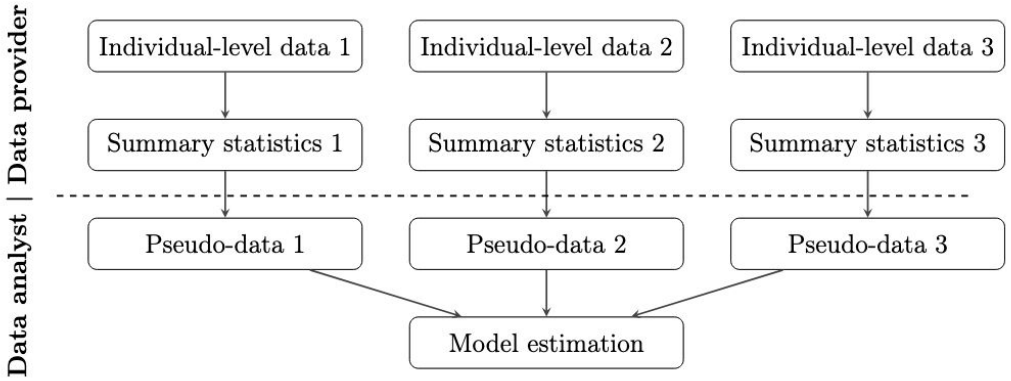


Figure 1: Proposed framework for a setup with three data providers. Each data provider prepares a summarized version of their data. The data analyst generates pseudo-data whose summary statistics match those supplied by the data providers. A global model can then be estimated from the pooled pseudo-data.

Actual **summary statistics**

≅

Pseudo-data **summary statistics**

Actual **log-likelihood**

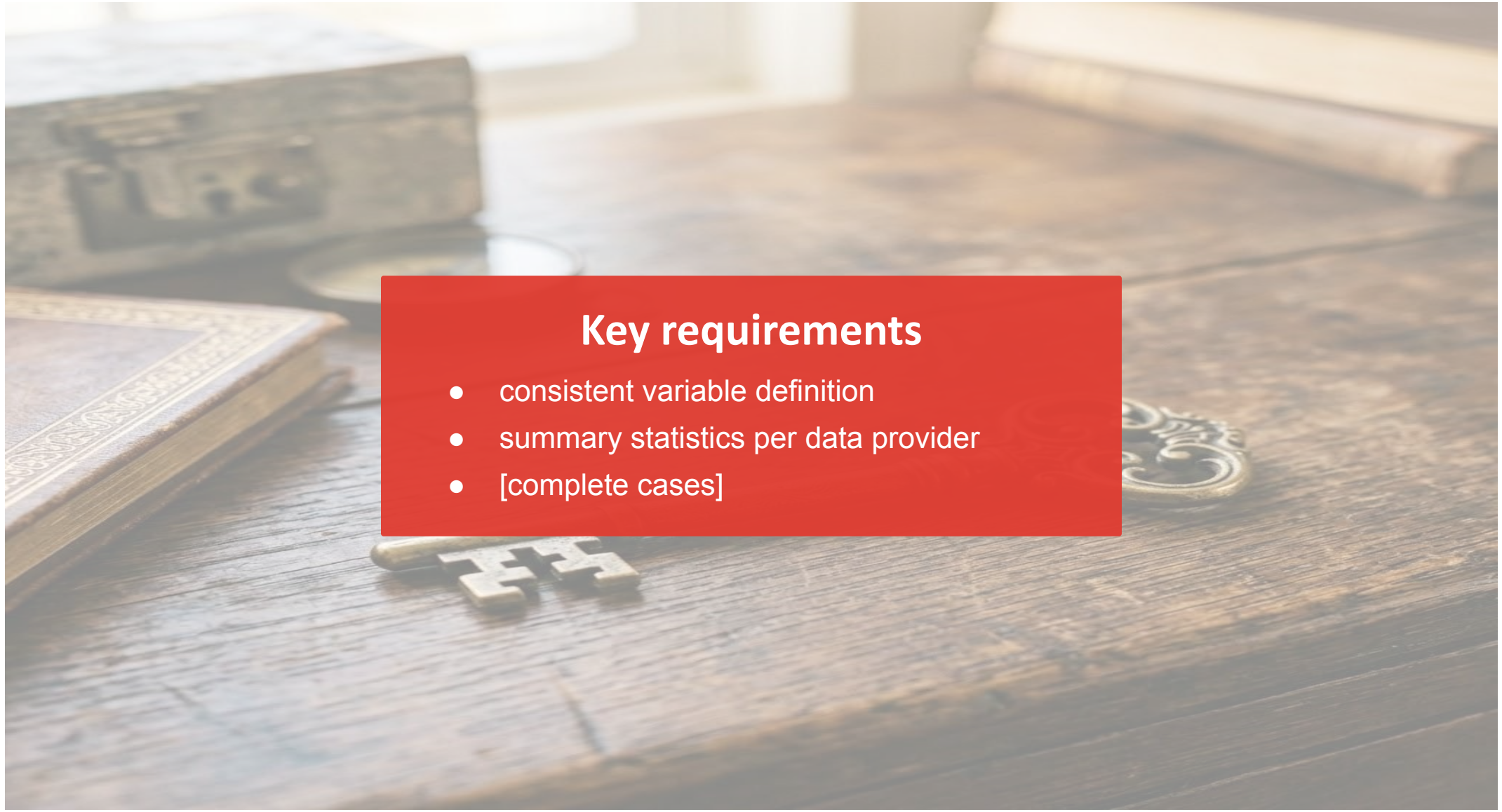
≅

Pseudo-data **log-likelihood**

Actual **inference**

≅

Pseudo-data **inference**


Key requirements
consistent variable definition
summary statistics per data provider
[complete cases]

Limitations

- x Exploratory data analysis
- x Outlier detection
- x Residuals analysis (model diagnostics)
- x Missing data imputation

Research potentials

Generalizable to other statistical models (e.g. linear geostatistical models)

Other correlation structures and random effects distributions

P(θ|D)

May be applicable to Bayesian models

Handling missing data

Final takeaway

To achieve precise inference, we do not need individual patient data;

we only need the information they contain.

M. A. A. Limpoco, C. Faes, and N. Hens. Linear mixed modeling of federated data when only the mean, covariance, and sample size are available. *Statistics in Medicine*, 44(1-2):e10300, 2025. doi: https://doi.org/10.1002/sim.10300.

M. A. A. Limpoco, C. Faes, and N. Hens. Federated mixed effects logistic regression based on one-time shared summary statistics. *Biometrical Journal*, 67(5):e70080, 2025. doi: https://doi.org/10.1002/bimj.70080.

M. A. A. Limpoco, C. Faes, and N. Hens. Federated generalized linear mixed models based on one-time shared summary statistics. (To be submitted).