



**Karolinska
Institutet**

Federated analyses

technical, statistical and human challenges

Bénédicte Delcoigne, Statistician, PhD

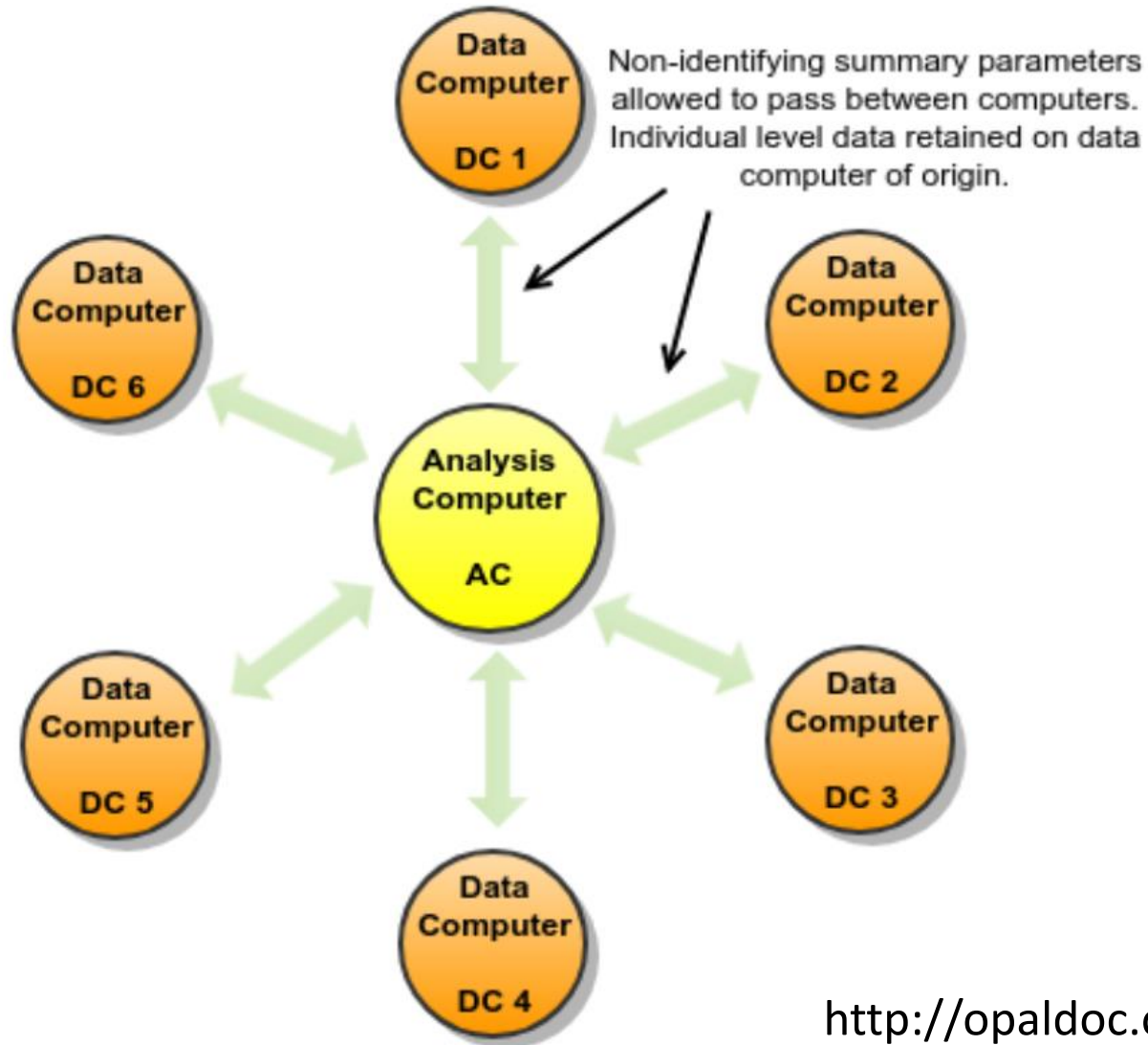
Department of Medicine (Solna), Unit of Clinical Epidemiology, Karolinska Institutet

What is it?

- When statistical power is the main limitation*
 - => increase the amount of data
 - pooling similar data => ethico-legal constraints & difficulties/reluctance for sharing data
 - performing a meta-analysis => pooled estimates
(Study Level Meta-Analysis SLMA)
- Federated analysis: centralized analysis with individual-level data remaining on local servers
 - => equivalent to meta-analysis at individual level
(Individual Level Meta-Analysis ILMA)
- Several solutions for performing federated analyses
- DataSHIELD package in R

* Not only...

How DataSHIELD is functioning



AC: Analysis Computer
DC: Data Computer

How does a federated analysis work?

An example with linear model (because it is simple)

dataset						
	y		x0	x1	x2	x3
obs	Carbo- hydrate			age	weight	protein
1	33		1	33	100	14
2	40		1	47	92	15
3	37		1	49	135	18
4	27		1	35	144	12
5	30		1	46	140	15
6	43		1	52	101	15
7	34		1	62	95	14
8	48		1	23	101	17
9	30		1	32	98	15
10	38		1	42	105	14

Vector y

Matrix x

7 observations in Sweden

3 observations in Danmark

Linear model: data splitted horizontally

Linear regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \text{error term}$

$$b = x^T \cdot y \cdot (x^T \cdot x)^{-1}$$

Estimate for $\beta_0, \beta_1, \beta_2$ and β_3

Using the entire data set or separated datasets gives the same vector of estimates b , provided that:

$x^T \cdot x$ for dataset_1 and $x^T \cdot x$ for dataset_2
 $x^T \cdot y$ for dataset_1 and $x^T \cdot y$ for dataset_2

are summed up before going further
 are summed up before going further

dataset						
	y		x0	x1	x2	x3
obs	Carbo-hydrate			age	weight	protein
1	33		1	33	100	14
2	40		1	47	92	15
3	37		1	49	135	18
4	27		1	35	144	12
5	30		1	46	140	15
6	43		1	52	101	15
7	34		1	62	95	14
8	48		1	23	101	17
9	30		1	32	98	15
10	38		1	42	105	14

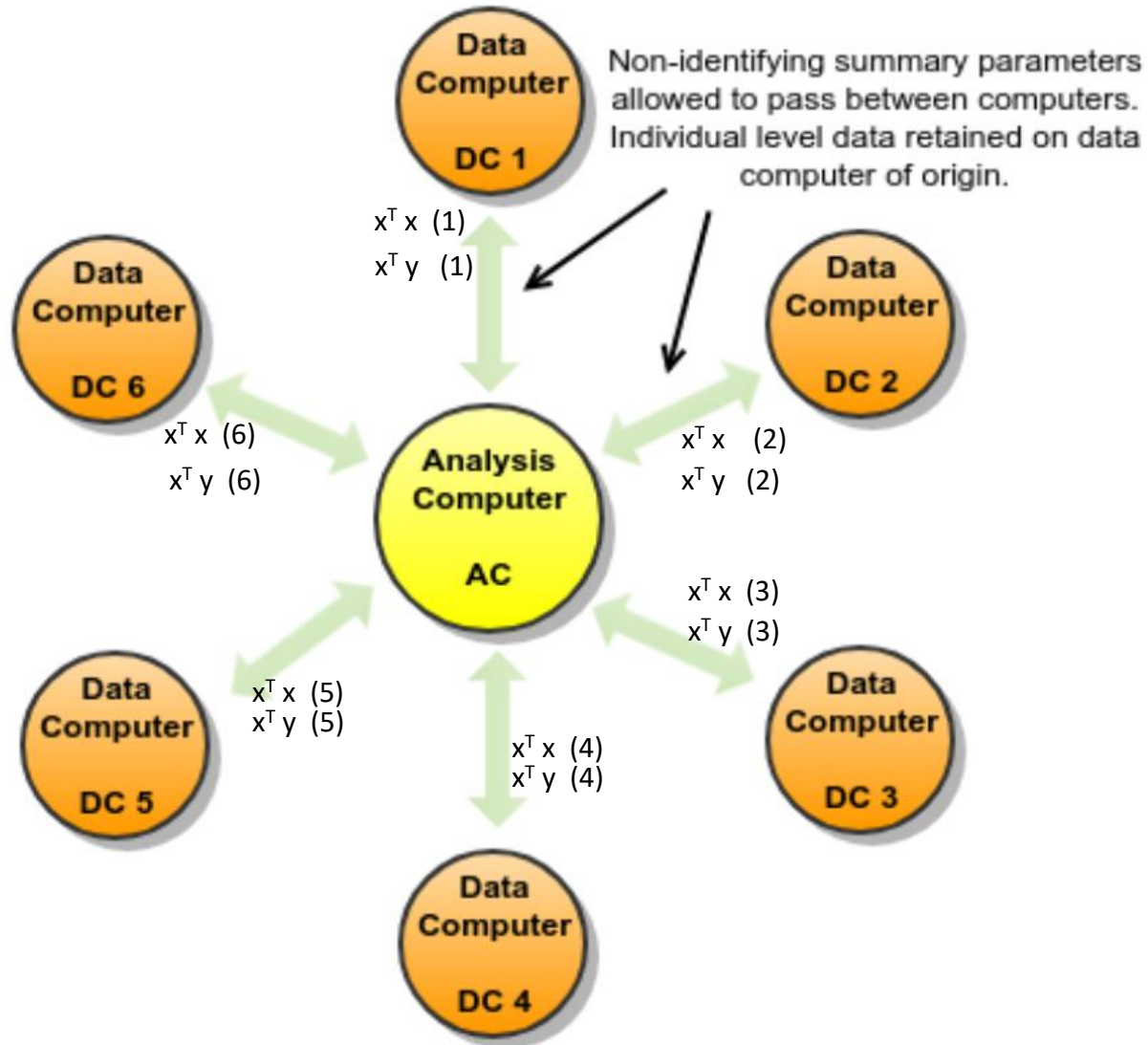
Dataset_1:

dataset						
	y		x0	x1	x2	x3
obs	Carbo-hydrate			age	weight	protein
1	33		1	33	100	14
2	40		1	47	92	15
3	37		1	49	135	18
4	27		1	35	144	12
5	30		1	46	140	15
6	43		1	52	101	15
7	34		1	62	95	14

Dataset_2:

dataset						
	y		x0	x1	x2	x3
obs	Carbo-hydrate			age	weight	protein
8	48		1	23	101	17
9	30		1	32	98	15
10	38		1	42	105	14

With DataSHIELD



AC: Analysis Computer
DC: Data Computer

Challenges

- Technical challenges: a good IT team is needed
- Statistical challenges: the choice of analysis
- What is doable right now with available tools in R:
 - DataSHIELD: Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual level Databases
 - distcomp: Computations over Distributed Data without Aggregation

Statistical challenges: DataSHIELD and GLM

Exponential family regression model

- Linear models: $E(Y_i) = \mu_i = \mathbf{x}_i^T \beta$ with $Y_i \sim N(\mu_i, \sigma^2)$
- Generalisation to non-linear function: link function $g(\mu_i) = \mathbf{x}_i^T \beta$

- General expression of the log likelihood of an outcome Y_i :

$$\text{Log } L(\theta_i, \varphi) = \frac{Y_i \theta_i - A(\theta_i)}{\varphi} + c(y_i, \varphi) \quad \text{with} \quad \begin{aligned} E(Y_i) &= A'(\theta_i) \\ \text{var}(Y_i) &= \varphi A''(\theta_i) \end{aligned}$$

- $A(\theta_i)$ chosen among standard models:

$$\begin{aligned} A(\theta_i) &= \frac{\theta_i^2}{2} && \text{Normal model} \\ A(\theta_i) &= \exp(\theta_i) && \text{Poisson model} \end{aligned}$$

DataSHIELD and GLM (continued)

- Iterative Weighted Least Squares (IWLS) for estimating β

Starting with β^0 the updating formula is:

$$\beta^1 = \beta^0 + \{ I(\beta^0) \}^{-1} S(\beta^0)$$

with:

$\{ I(\beta^0) \}^{-1}$ the inverse of the Fisher information matrix, the variance-covariance matrix of the parameter estimates

$S(\beta^0)$ the score function

DataSHIELD and GLM (continued)

Step 1: Transmission AC -> DC

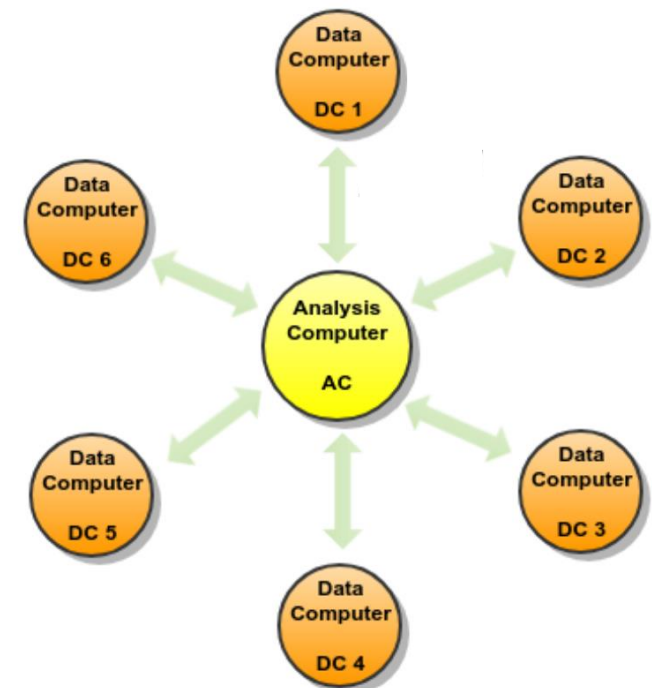
lines of code to run the statistical model (1st iteration)
and obtain necessary elements to continue
(in linear model example: $x^T y$ and $(x^T \cdot x)$)

Step 2: Transmission DC -> AC

the computed vector and matrix
(in linear model: $x^T y$ and $(x^T x)$)

If GLM, fitted iteratively with IWLS,
there are several steps, going back and forth
until convergence

Final step : AC compute the estimates



Statistical challenges

- Why Cox analysis is not doable?
Because of the Cox partial likelihood:

$$L(\beta) = \prod_{t_i} \frac{\exp[\beta X_i]}{\sum_{k \in R_i} \exp[\beta X_k]}$$

- ? mourning the study we dreamed of ... and being flexible
- As GLM and Lexis are available in DataSHIELD, a Poisson regression can be performed instead of a Cox.
- A Cox stratified analysis is doable with distcomp R-package.

Linear model: data splitted **vertically**

Linear regression $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \text{error term}$

$$b = \underbrace{x^T \cdot y}_{\text{red}} \cdot \underbrace{(x^T \cdot x)^{-1}}_{\text{blue}}$$

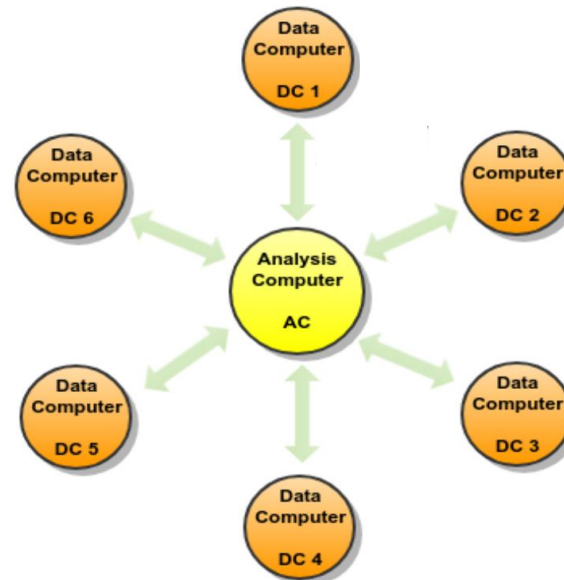
$x^T \cdot x$ as before _ no problem

$x^T \cdot y$???

Dataset_1:

Dataset_2:

dataset			x0	x1	x2	x3
	y			age	weight	protein
obs	Carbo- hydrate					
1	33		1	33	100	14
2	40		1	47	92	15
3	37		1	49	135	18
4	27		1	35	144	12
5	30		1	46	140	15
6	43		1	52	101	15
7	34		1	62	95	14
8	48		1	23	101	17
9	30		1	32	98	15
10	38		1	42	105	14



Other challenges

- Data harmonisation:

- Step 1:

- outcome definition: icd codes

- selection of study individuals

- which variables are available? How/when measured?

- Do they mean the same thing?

- Ex: disease duration, drug name, drug start date, ...

- Step 2:

- same names, same formats/coding, same degree of precision, ...

- Ex: coding of sex

dataset						
	y		x0	x1	x2	x3
obs	Carbo- hydrate			age	weight	protein
1	33		1	33	100	14
2	40		1	47	92	15
3	37		1	49	135	18
4	27		1	35	144	12
5	30		1	46	140	15
6	43		1	52	101	15
7	34		1	62	95	14
8	48		1	23	101	17
9	30		1	32	98	15
10	38		1	42	105	14

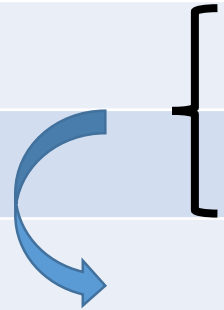

In our team

- Goal: to perform a pilot study using a federated analysis approach
- Swedish and Danish registers data
- Research question
“risk of neurological side effects of Tumor Necrosis Factor alpha-inhibitors (TNFi) among patients with arthritis”
- Study design:
 - cohort of arthritis patients
 - followed from registration (≥ 2001) until event/death/end 2016
 - patients enter the cohort either unexposed or exposed to TNFi
- Statistical analysis: Cox with time-varying exposure

Step we are involved right now

- Simulation of a cohort that we can analyse both with time-varying (exposure and covariates) Cox and Poisson with time-varying exposure and covariates
- in R
- Goal: compare estimates obtain with different statistical approaches and test DataSHIELD
- Why simulation? No constraint in sharing

Preliminary results on 1 simulated cohort

Model	HR (95% Confidence Interval)
Cox	2.05 (1.50 - 2.79)
Cox _ country 1	 1.90 (1.30 - 2.79)
Cox _ country 2	
Cox _ meta-analysis	
Poisson	2.05 (1.50 - 2.79)
Poisson _ country 1	 1.90 (1.30 - 2.78)
Poisson _ country 2	



**Karolinska
Institutet**

Thanks for your attention

References:

Wolfson M. et al. *International Journal of Epidemiology* 2010;39:1372–1382
Jones E.M. Et al. *Norsk Epidemiologi* 2012; 21 (2): 231-239