DataSHIELD

What is it and when is it useful?

Andrei S. Morgan, FRCPCH, MSc. PhD

andrei.morgan@inserm.fr

Institute National de la Sante et de la Recherche Médicale, Paris, France

LSHTM R users group, April 2020

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 733280.

Disclosure and acknowledgements

This presentation derives from work performed as a sub-project of the RECAP Preterm collaboration – `https://recap-preterm.eu`

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

Disclosure and acknowledgements

This presentation derives from work performed as a sub-project of the RECAP Preterm collaboration – `https://recap-preterm.eu`

Big thanks to...

RECAP Preterm:

- **Helen Collins**, Leicester (data harmonisation)
- **Gonçalo Gonçalves**, INESC TEC, Portugal (infrastructure)
- **Jennifer Zeitlin**, INSERM, France (RECAP WP7 Project Lead)
- **Marina Cuttini**, OPBG, Italy (ACTION cohort PI)

DataSHIELD:

- **Tom Bishop**, Cambridge
- **Demetris Avraam**, Newcastle (statistics)
- **Alex Westerberg**, Newcastle (user experience)
- **Paul Burton**, Newcastle (DataSHIELD PI)

All errors are my own, and I have no particular skills in any of this...

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 733280.

Disclosure and acknowledgements

This presentation derives from work performed as a sub-project of the RECAP Preterm collaboration – `https://recap-preterm.eu`

Big thanks to...

RECAP Preterm:

- **Helen Collins**, Leicester (data harmonisation)
- **Gonçalo Gonçalves**, INESC TEC, Portugal (infrastructure)
- **Jennifer Zeitlin**, INSERM, France (RECAP WP7 Project Lead)
- **Marina Cuttini**, OPBG, Italy (ACTION cohort PI)

DataSHIELD:

- **Tom Bishop**, Cambridge
- **Demetris Avraam**, Newcastle (statistics)
- **Alex Westerberg**, Newcastle (user experience)
- **Paul Burton**, Newcastle (DataSHIELD PI)

All errors are my own, and I have no particular skills in any of this...

Conflict of interest: Chair of the DataSHIELD Advisory Board.

(I'm also a neonatologist and epidemiologist)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

Background

Context

- Many studies are now integrating data collected in different locations
- Individual Patient Data (IPD) - permits “real” comparison
- Data transfer is complicated
  - Legal issues
  - Ethical issues
  - Time-consuming

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 733280.

Background

Context

- Many studies are now integrating data collected in different locations
- Individual Patient Data (IPD) - permits “real” comparison
- Data transfer is complicated
  - Legal issues
  - Ethical issues
  - Time-consuming

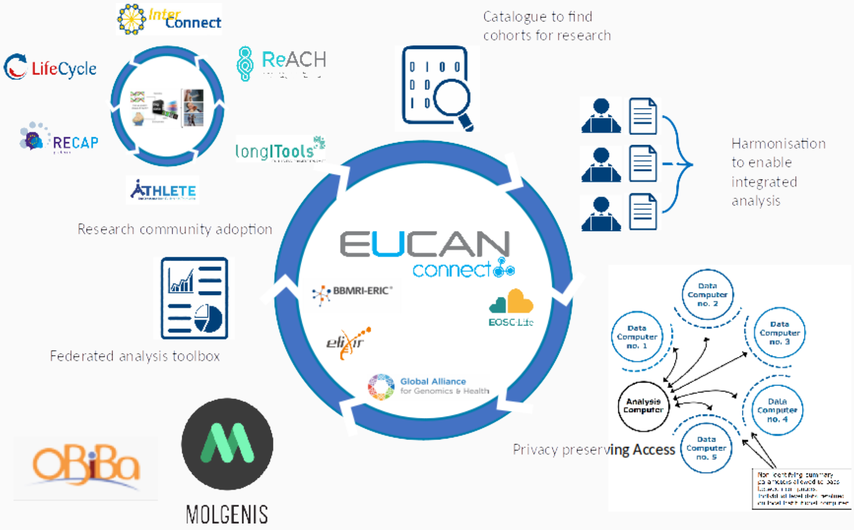
DataSHIELD is a potential solution to these problems

DataSHIELD
Secure Bioscience Collaboration

https://www.datashield.org

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 733280.

Making cohort research FAIR



Catalogue to find cohorts for research

Harmonisation to enable integrated analysis

Privacy preserving Access

Federated analysis toolbox

Research community adoption

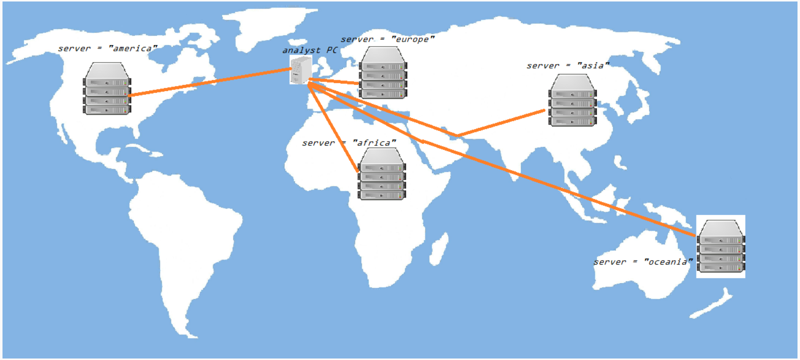
Non identifying summary parameters allowed to pass between computers. Individual level data remains on local health board computer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

What is DataSHIELD?

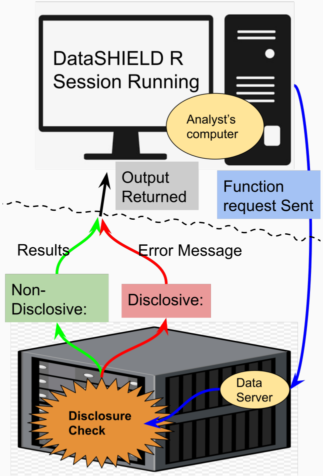
Simply put, DataSHIELD is:

- A *privacy-preserving, federated analysis* software
- It can analyse sensitive data, at an individual level, remotely (without direct access)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

How does DataSHIELD work?



- Functions are sent to data servers (nodes)
- Server(s) run functions
- Functions contain built-in disclosure checks – if:
  - *non-disclosive*: returns results
  - *disclosive*: returns error message
- Analyst computer collates results

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

Study process

- Identify research question
- Agreement from cohorts for participation
- Identify variables required for analysis
- Variable harmonisation
- Provide access to nodes
- Carry out statistical analyses
- Get results
- Publish!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

Impact of sex on survival to discharge

Study question:

Does perinatal mortality among extremely preterm births differ by (fetal) sex?

(Extreme preterm birth: birth before 27 weeks of gestation)

Hypothesis: The sex ratio at birth of extreme preterms is biased towards males (i.e. more live born babies are male).

Hypothesis: Survival to hospital discharge of live born babies is higher in females.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

Impact of sex on survival to discharge

Study question:

Does perinatal mortality among extremely preterm births differ by (fetal) sex?

(Extreme preterm birth: birth before 27 weeks of gestation)

Hypothesis: The sex ratio at birth of extreme preterms is biased towards males (i.e. more live born babies are male).

Hypothesis: Survival to hospital discharge of live born babies is higher in females.

- Agreement from cohorts
- Identified variables from cohort dictionaries / *a priori* reasoning
- Variable harmonisation — by data managers on nodes (**see:** `https://platform.recap-preterm.eu`)
- Access to nodes granted to analysts by data managers
- Statistical analysis ...

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.

Links and outputs

Links:

- `https://www.datashield.org` – DataSHIELD 'R' software
- `https://www.obiba.org` – Obiba stack (data curation and harmonisation)
- `https://platform.recap-preterm.eu` – RECAP Preterm data analysis platform

Papers:

- Marcon et al, **Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD**, PLOS Computational Biology (2021) doi: 10.1371/journal.pcbi.1008880
- Wilson et al, **DataSHIELD – New Directions and Dimensions**, *Data Science Journal* (2017), doi: 10.5334/dsj-2017-021
- Doiron et al, **Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination**, *International Journal of Epidemiology* (2017), doi: 10.1093/ije/dyx180
- Pearce et al, **Associations of Total Legume, Pulse, and Soy Consumption with Incident Type 2 Diabetes: Federated Meta-Analysis of 27 Studies from Diverse World Regions**, *Journal of Nutrition* (2021), doi: 10.1093/jn/nxaa447

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280.