DISCERN

2025 annual progress report

International Agency for Research on Cancer
World Health Organization

Vivian Viallon
Ali Farnudi

Over view

- Bird’s eye view of the project
- Pre-processing of untargeted metabolomics data
- Construction and evaluation of high-dimensional signatures of cancer risk in "untypical" case-cohort studies
- Federated analysis of cohort (and case-cohort) data
- Publishing of a public website for all developed tools on a collaborative platform

Somalogic data

Covariates
age, gender, etc

event

Metabolic data
LC-MS + GC-MS

Centralised at IARC
Case cohort + case series

Pre-processing of untargeted metabolomics data

The “IARC pipeline” for IARC LC-MS data

Pipelines for univariate multivariate analysis

Model evaluation

Simulation

- F1 score
- C-index
- Time dependant AUC

De-centralised data

Federated analysis

Somalogic data

Covariates
age, gender, etc

event

Metabolic data
LC-MS + GCMS

Pipelines will be
publically available on a
collaborative platform with
Documentation

Pre-processing of untargeted metabolomics data

The "IARC pipeline" for IARC LC-MS data

Centralised at IARC
Case cohort + case series

De-centralised data

Federated analysis

Pipelines for univariate multivariate analysis

Model evaluation

Simulation

- F1 score
- C-index
- Time dependant AUC

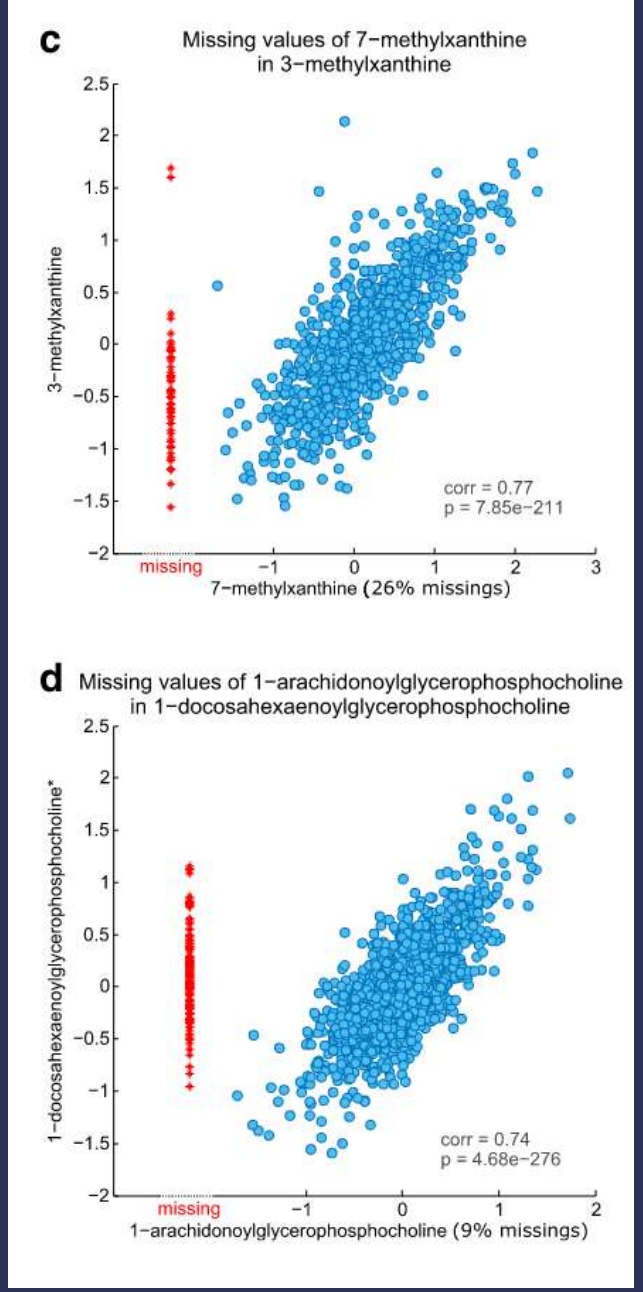
Pre-processing of untargeted metabolomics data

The “IARC pipeline” for IARC LC-MS data

- Input: data measured and “curated” by Pekka’s lab
- Pre-processing steps implemented in an R package
  1. Filtering of features with too many missing values
  2. Filtering of samples with too many missing values, and outlier samples
  3. Imputation of missing data
     - LCMD + RF (to be added)
  4. Normalization
     - plate correction
     - intensity drift correction (to be added)
  5. Clustering based on RT and correlations (optional; to be added)
- Output: data ready for statistical analyses
  - Various versions can be produced, for sensitivity analyses

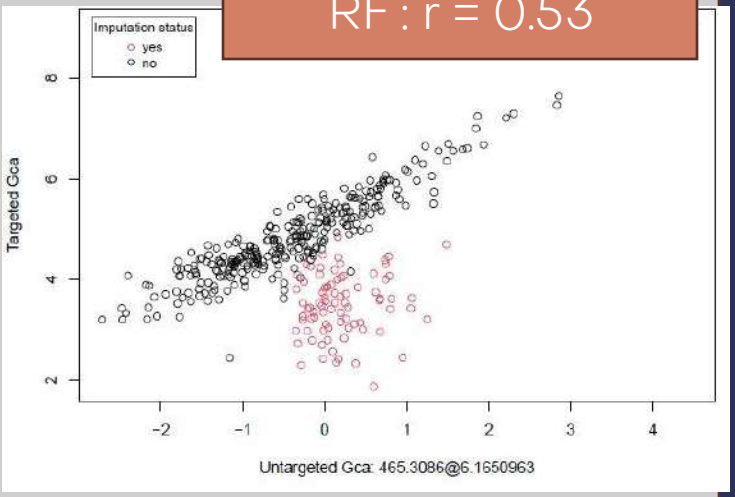
Imputation of missing values

- Pekka and people from the lab: “Missing values are generally <LOD” : use LOD/2 or LCMD
- Do et al. (Krumsiek’s lab). “Not always”: use KNN or RF
- Hybrid approach:
  - RF if it performs really well
  - LCMD otherwise



Comparison with targeted metabolomics (bile acids) data in EPIC

RF : r = 0.53



Hybrid r = 0.90

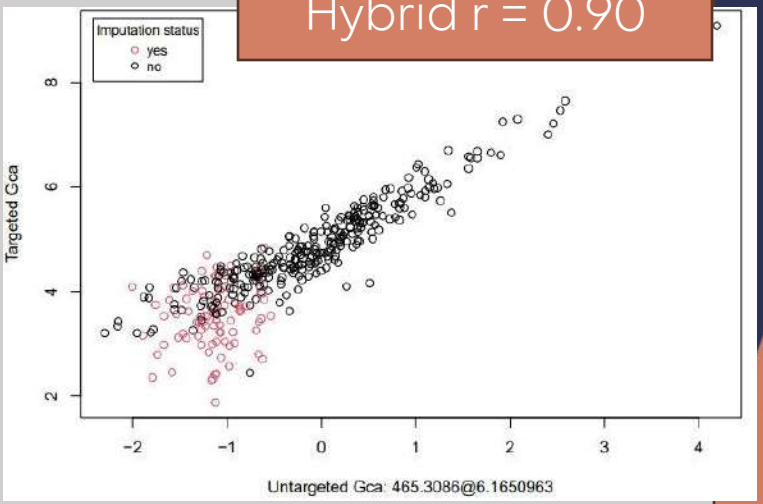
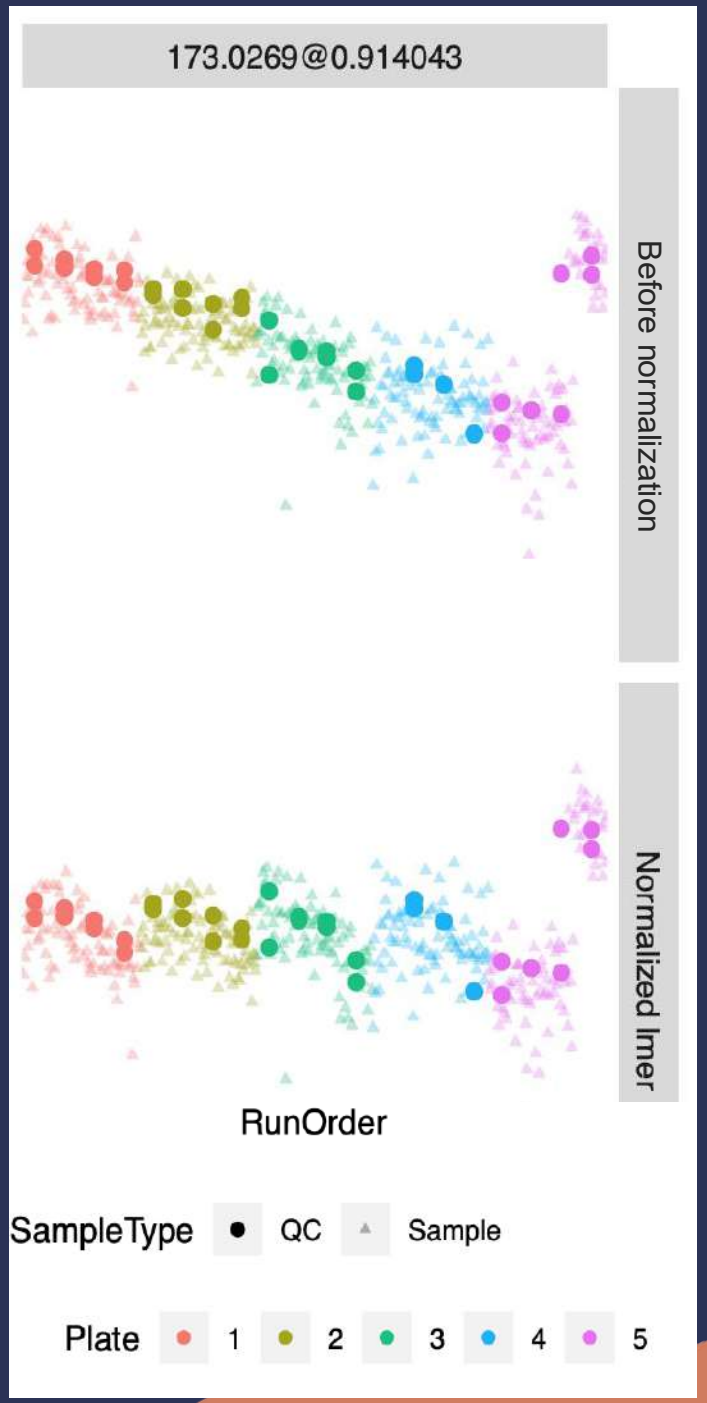


Plate correction… is not always enough

- Inspection of QCs and/or run order is useful
- Correcting for plate effects is not good enough for some features

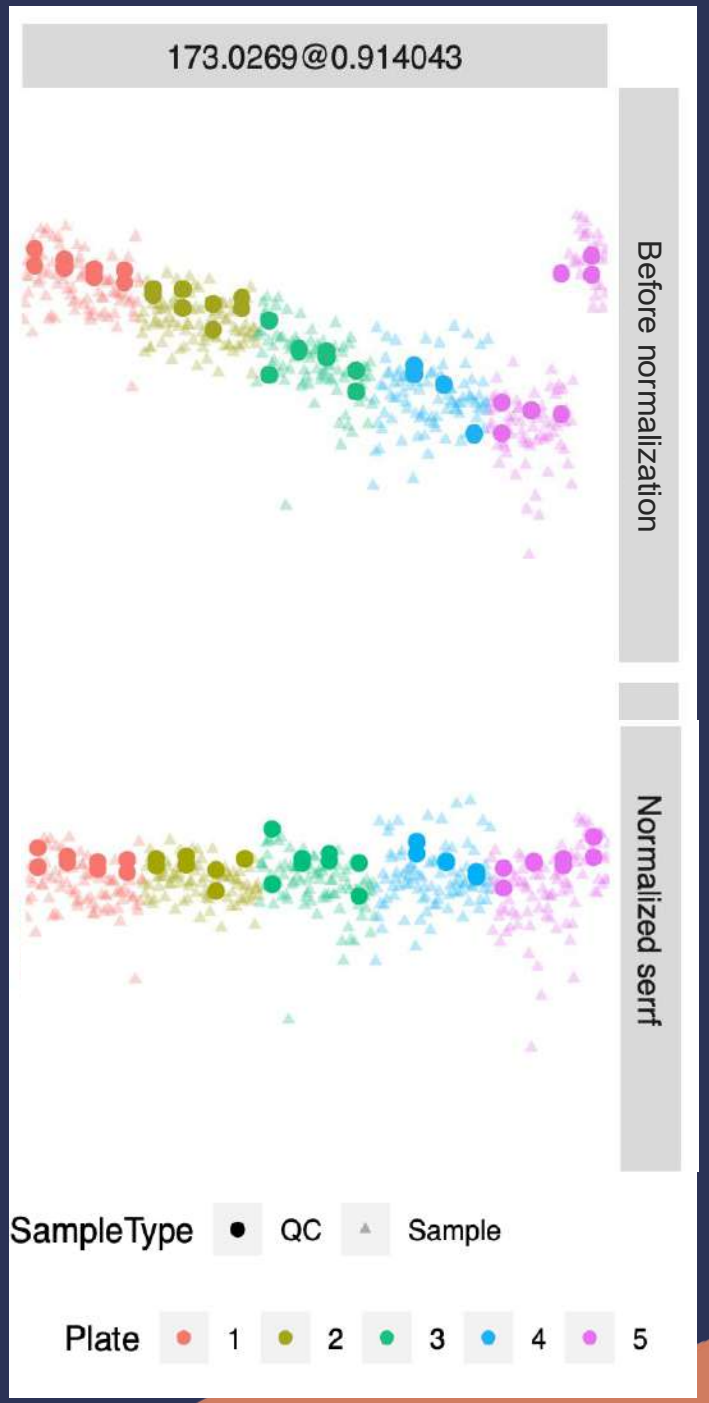


SERRF to correct for "intensity drift"

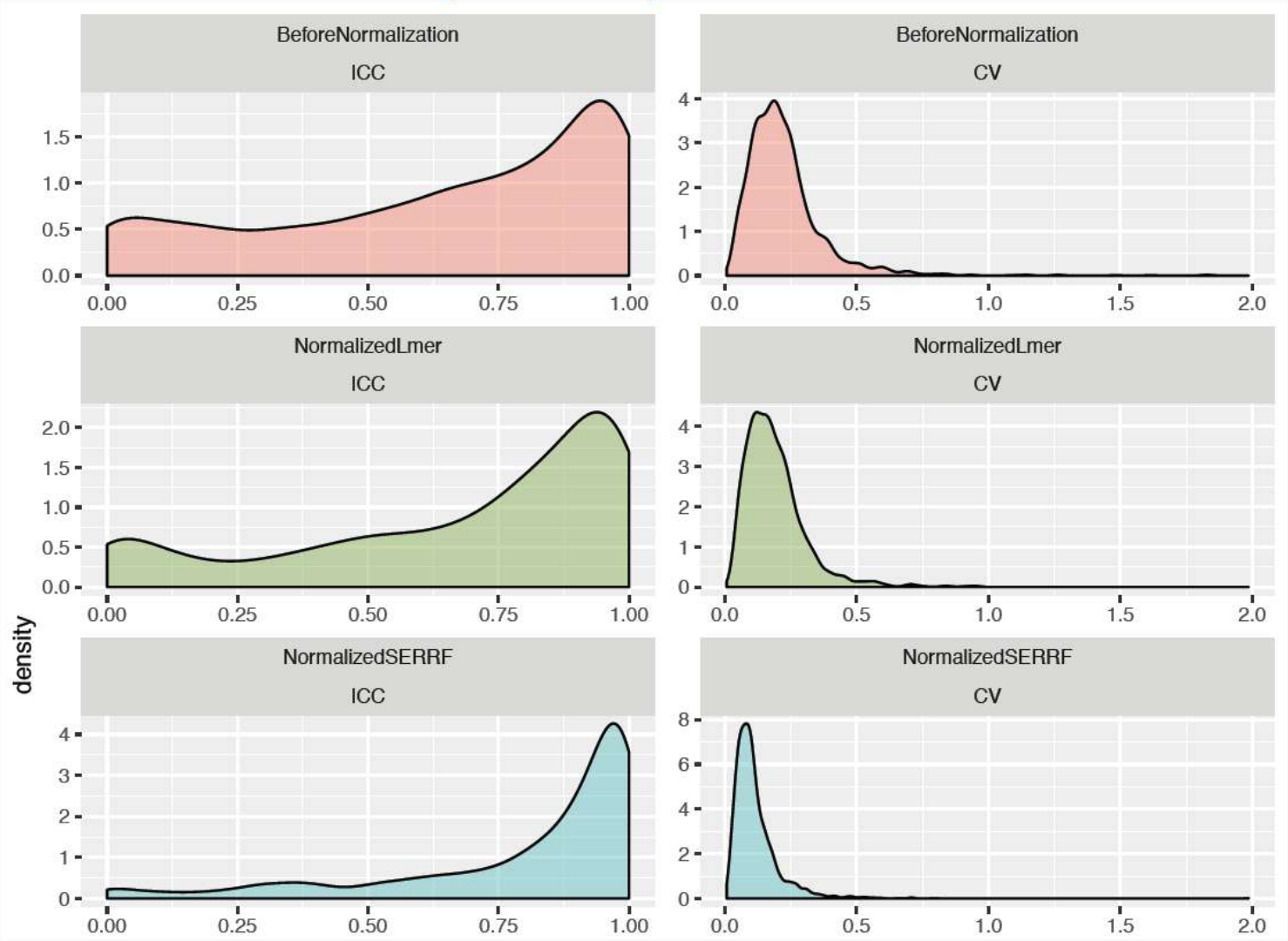
- [Fan et al. (2019). Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. Anal. Chem.]

- **Try different approaches and check consistency of the end results**

- **Visual inspection is helpful**

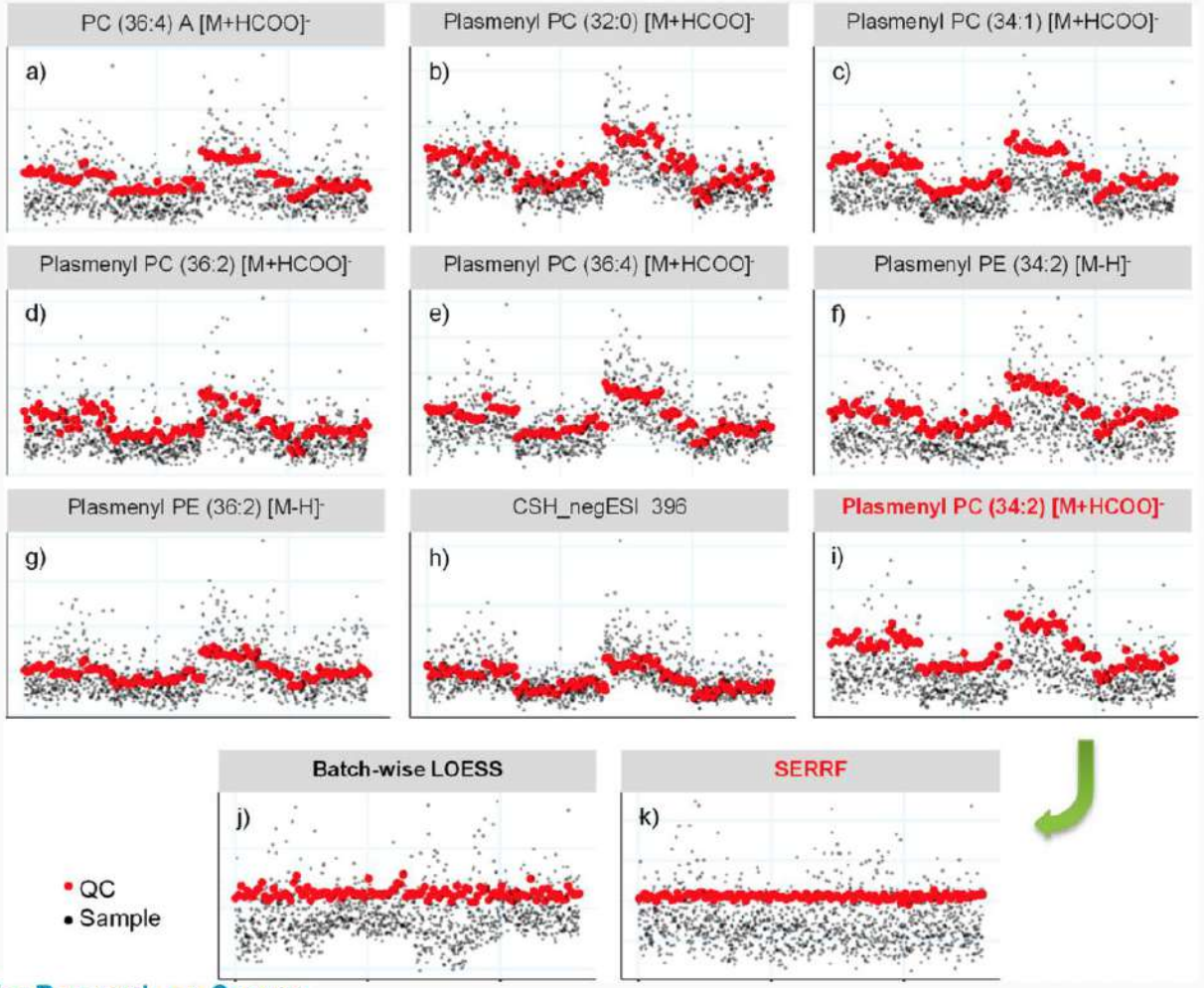


ICC and CV


Distribution of ICC_j and CV_j, $j = 1, \ldots, p$ (POS mode)
BeforeNormalization
ICC
BeforeNormalization
CV
NormalizedLmer
ICC
NormalizedLmer
CV
NormalizedSERRF
ICC
NormalizedSERRF
CV
density

SERRF

[Fan et al. (2019). *Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data*. Anal. Chem.]

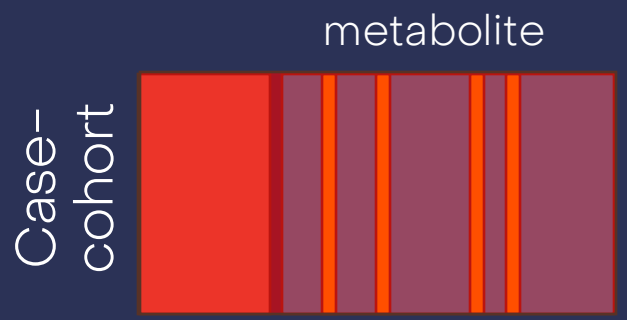


Next steps

- Finalize the package with “IARC”-methods
- Integrate alternative/complementary approaches (eg., from Marc and EXPANSE?)
- Illustrate the use of the package on some of the DISCERN metabolomics data (Rmarkdown)
- Extension to Recetox GC-MS data

Construction and evaluation of high-dimensional signatures of cancer risk in "untypical" case-cohort studies

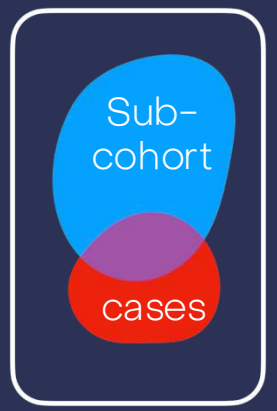
Key aims

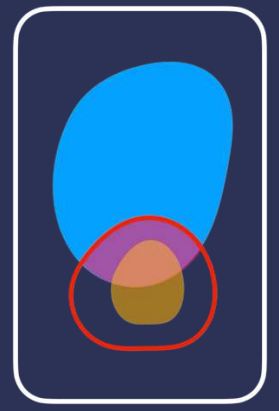


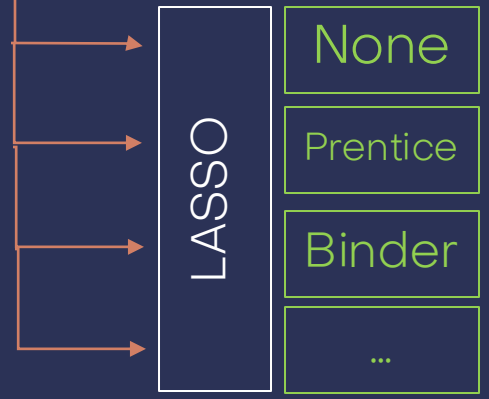
Identify associated metabolite

Model and predict + weights?

Simulate cohort data







Performance Evaluation

f1-score: selection

predictive power C-index

Time-dependant AUC
work in progress...

Cohort

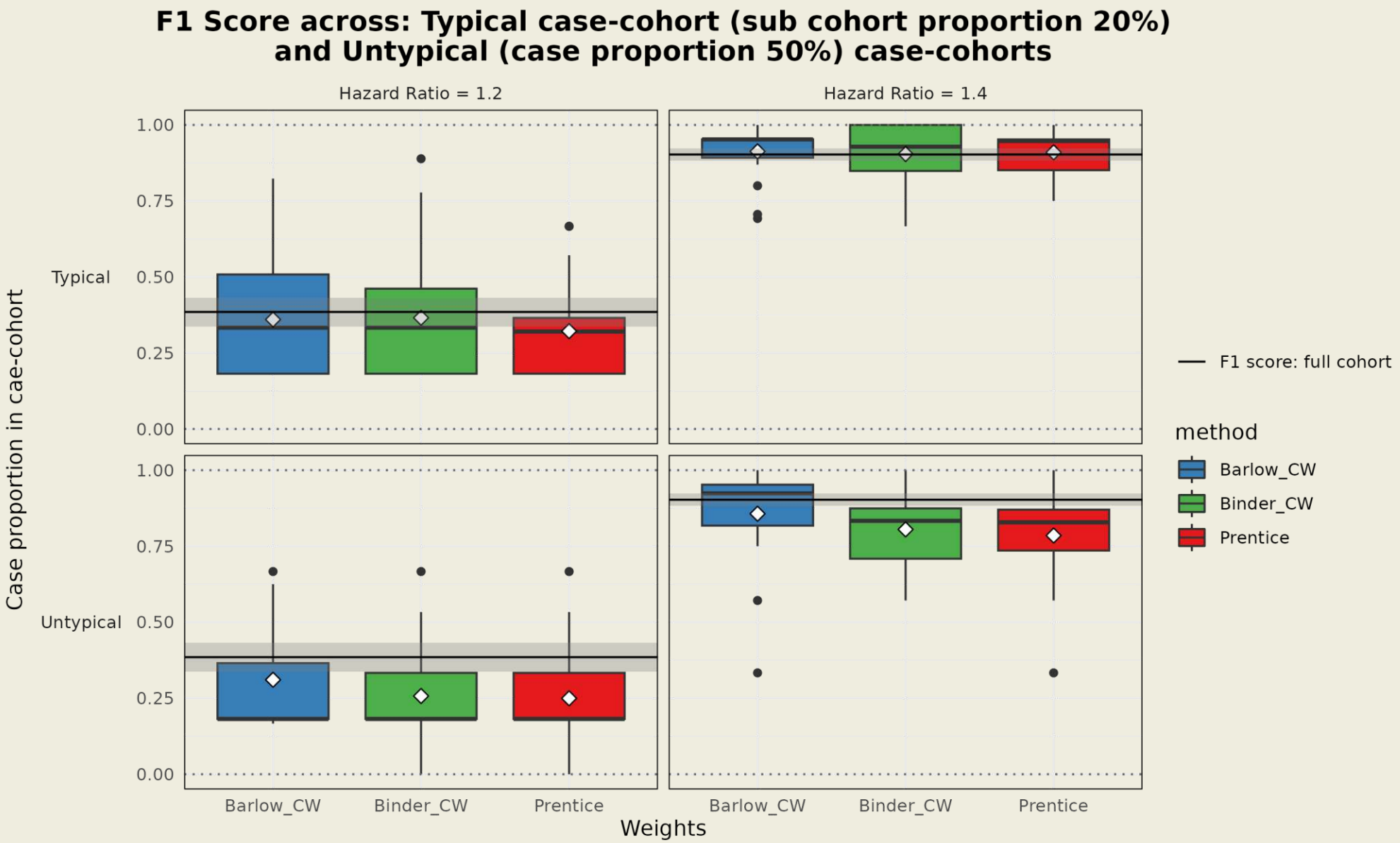
Case-cohort

None | Prentice | ...

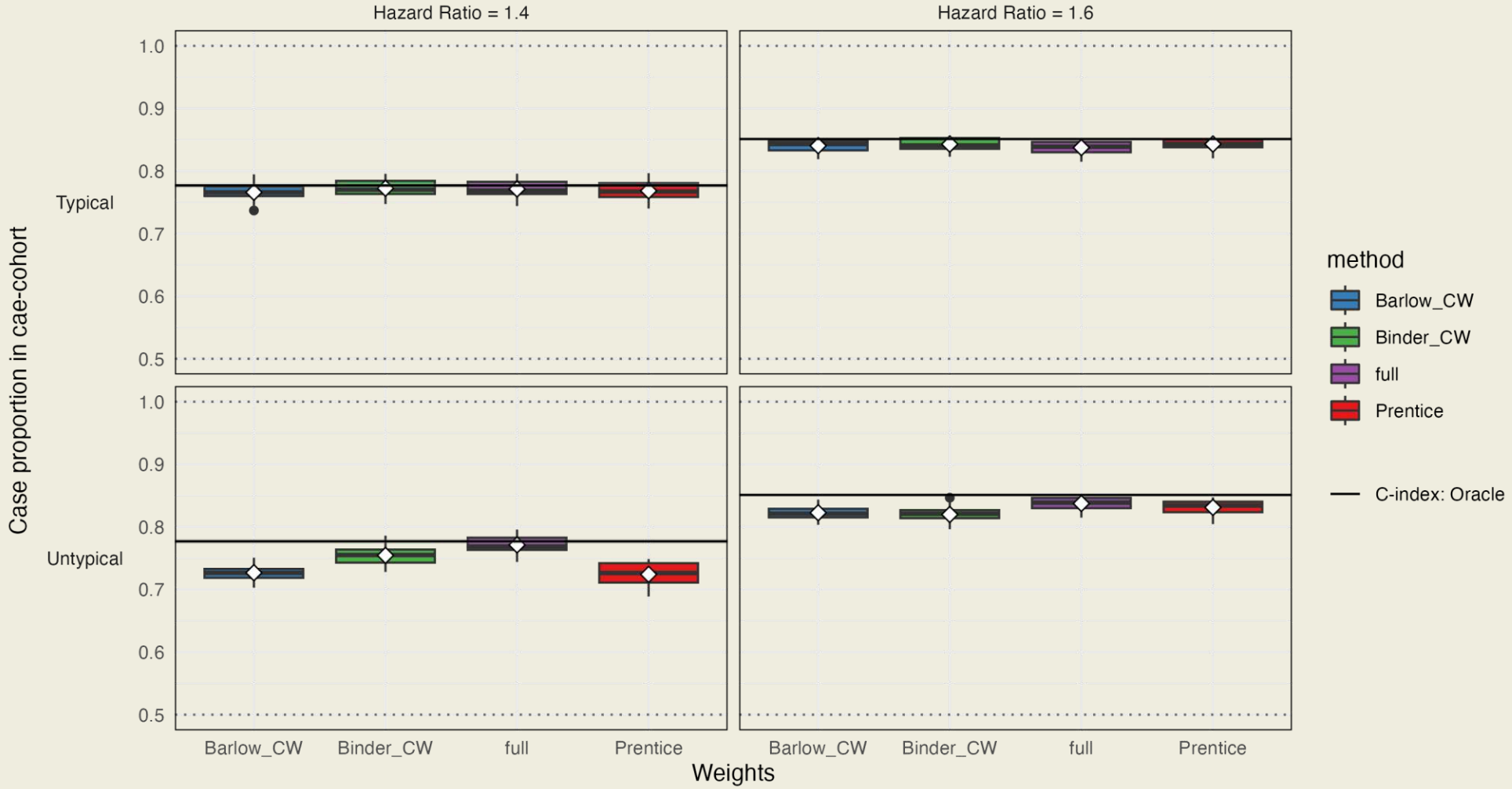
Conclusion

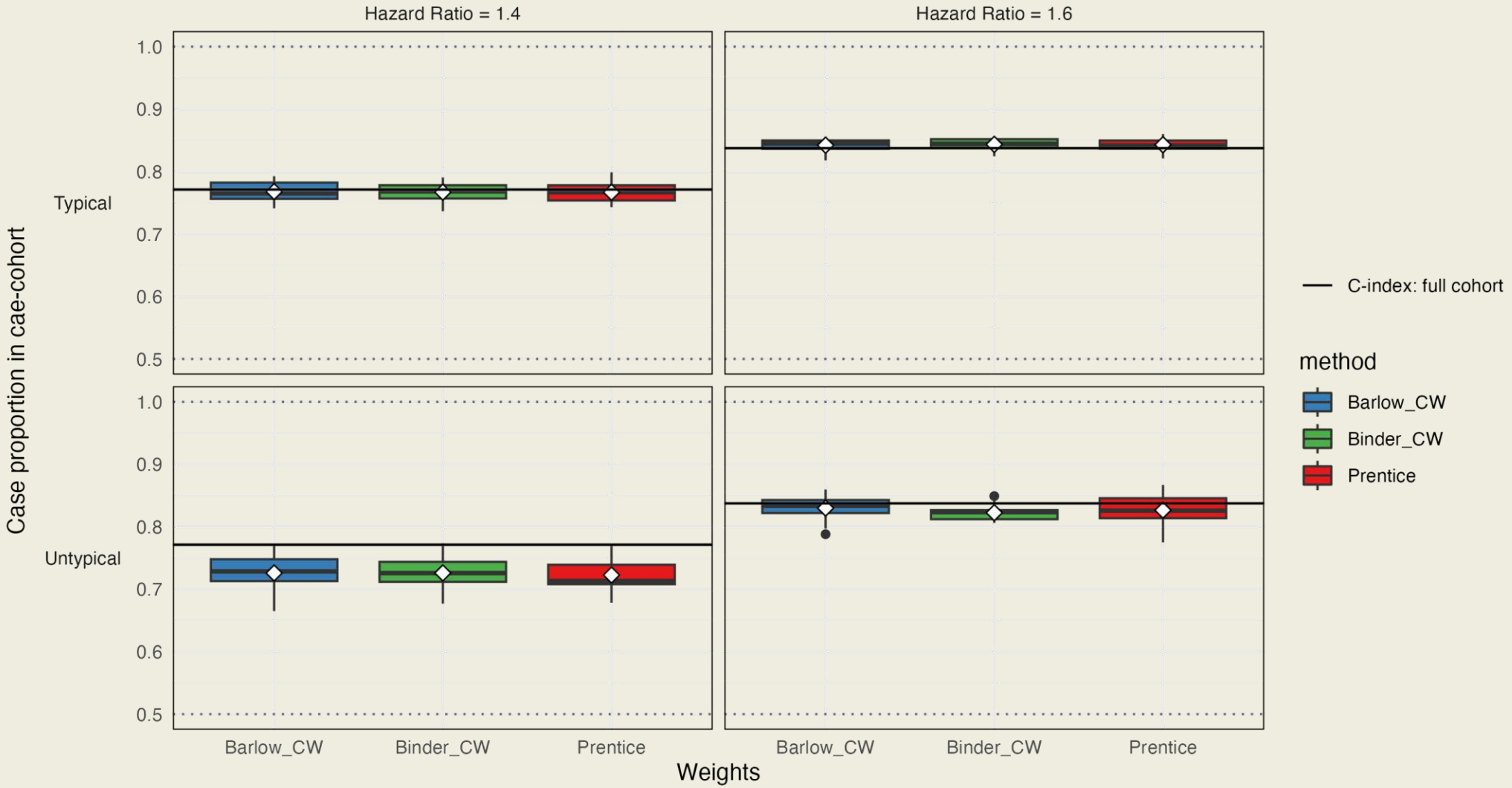
- Performance of weighted Lasso models:
  - variable selection: all versions perform similarly
  - Discrimination: Barlow weights perform slightly better

- Performance of the C-index estimate on case-cohort data
  - Analysis in progress


F1 Score across: Typical case-cohort (sub cohort proportion 20%) and Untypical (case proportion 50%) case-cohorts
Hazard Ratio = 1.2
Hazard Ratio = 1.4
Typical
Untypical
Case proportion in cae-cohort
1.00
0.75
0.50
0.25
0.00
Barlow_CW
Binder_CW
Prentice
Weights
F1 score: full cohort
method
Barlow_CW
Binder_CW
Prentice

C-index estimation of models with differnet weights across: Typical case-cohort (sub cohort proportion 20%) and Untypical (case proportion 50%) case-cohorts




Performance of the C-index estimation for lasso + Prentice, on case-cohort designs (sub cohort proportion 20%) and Untypical (case proportion 50%) case-cohorts
Hazard Ratio = 1.4
Hazard Ratio = 1.6
Typical
Untypical
Case proportion in cae-cohort
1.0
0.9
0.8
0.7
0.6
0.5
Barlow_CW
Binder_CW
Prentice
Weights
C-index: full cohort
method
Barlow_CW
Binder_CW
Prentice

Next steps

- Add time-dependent AUC analysis.

- Consider alternatives to the simple lasso
  - David's approach?
  - Lasso + stability selection (eg the sharp package)

- Deliver simulation pipeline, tools, and weight recommendations as an R package.

- Publish a comprehensive webpage for application of tools and reproducibility of all results for DISCERN partners

- Federated analysis of cohort (and case-cohort) data

Federated Computation

Meta analysis

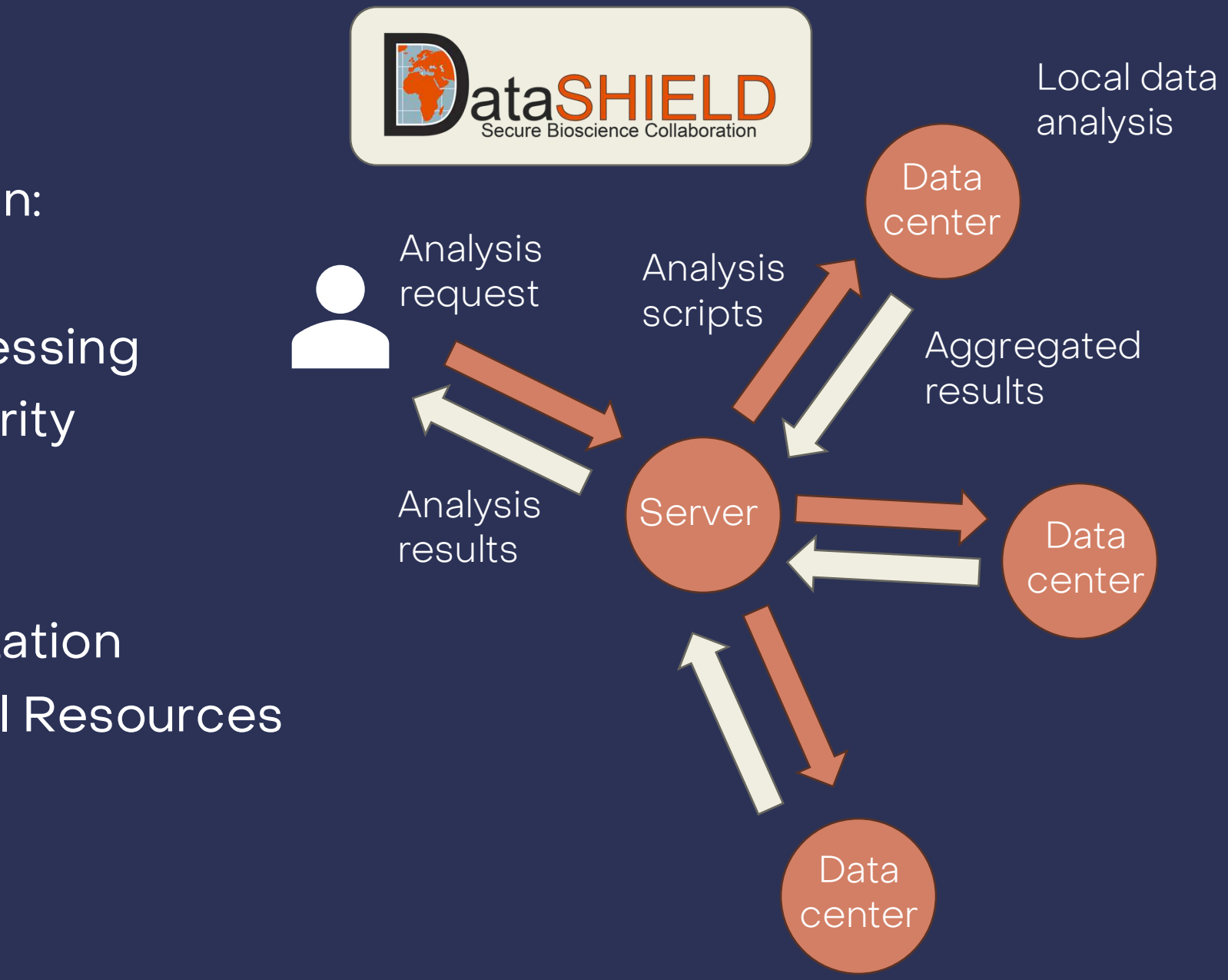
Federated analysis

Federated computation:

- Decentralised data
- Decentralised Processing
- Data Privacy & Security

Limitations:

- Complex Implementation
- Local Computational Resources



Federated Computation

Meta analysis

Federated analysis

Meta analysis

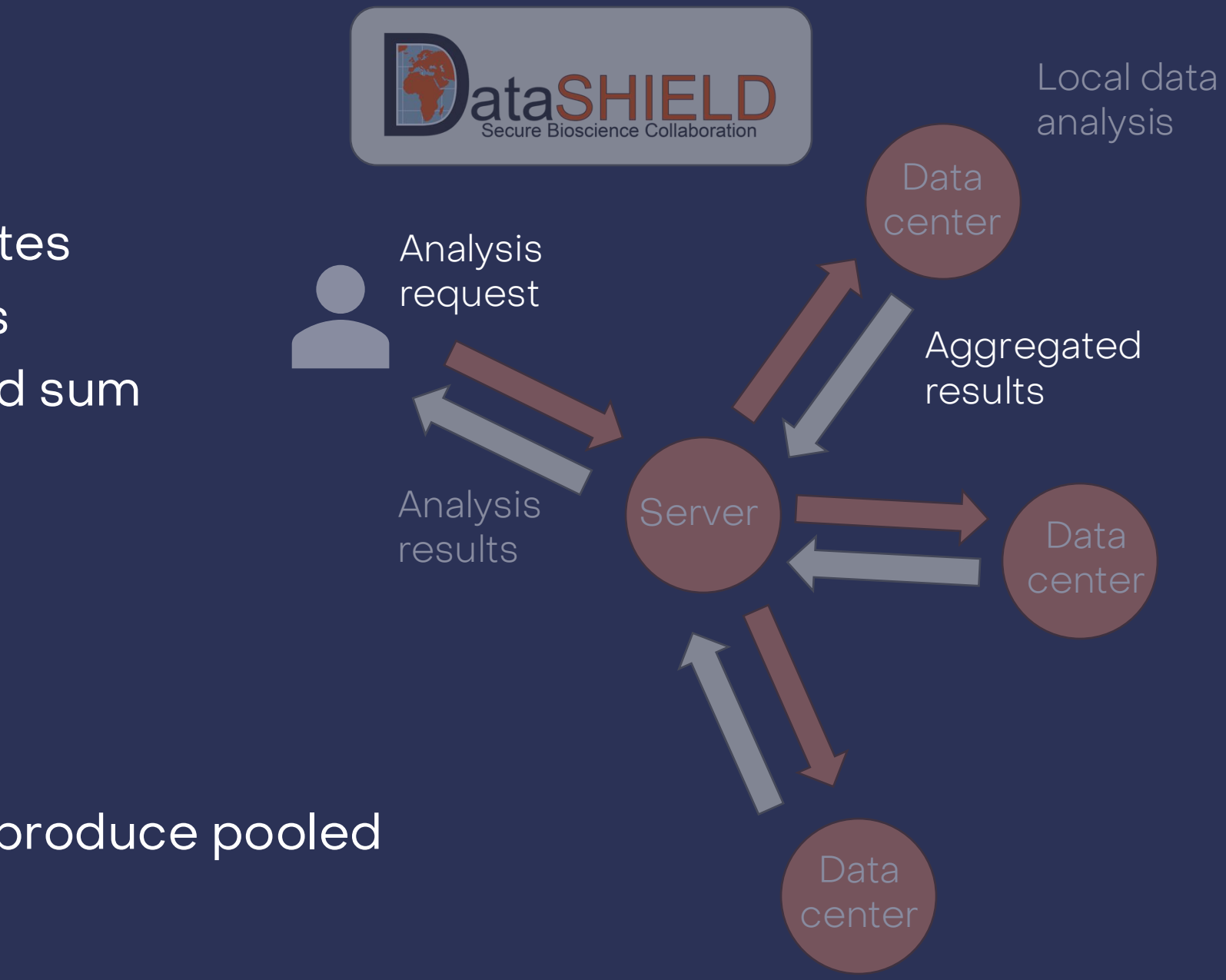
- Coefficient estimates
- Variance estimates
- Perform a weighted sum

Advantages:

- Simple

Limitations:

- Not intended to reproduce pooled estimates



Terminology

Federated Computation

Meta analysis

Federated analysis

Federated analysis

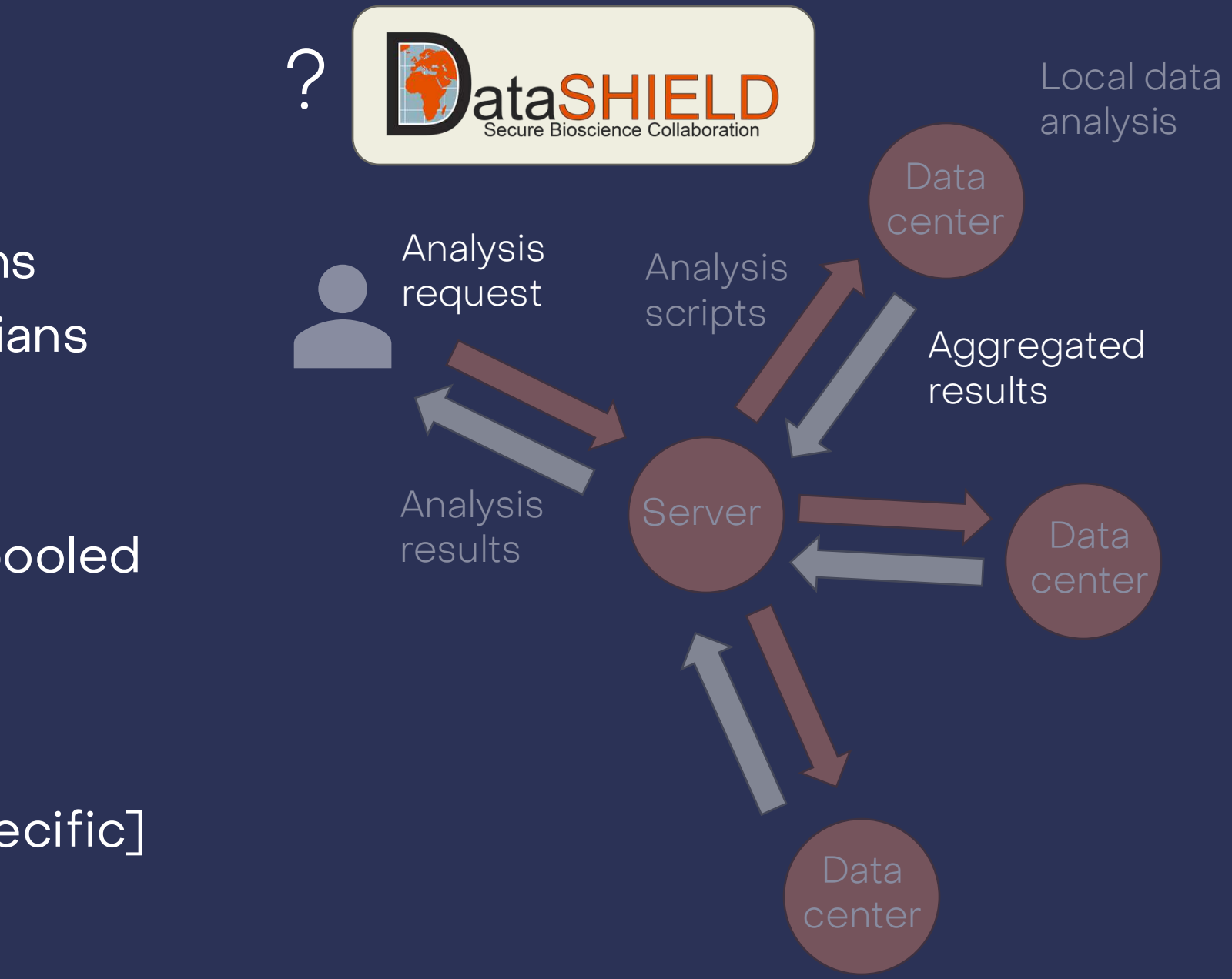
- Federated algorithms
- Gradients and Hessians

Advantages:

- Intended to mimic pooled estimates

Limitations:

- Complex [model specific]



Federated analysis
+
Somalogic data

EPIC -> Proteomics Data -> Protein concentrations ~7500

- Italy
- Spain
- The Netherlands
- United Kingdom

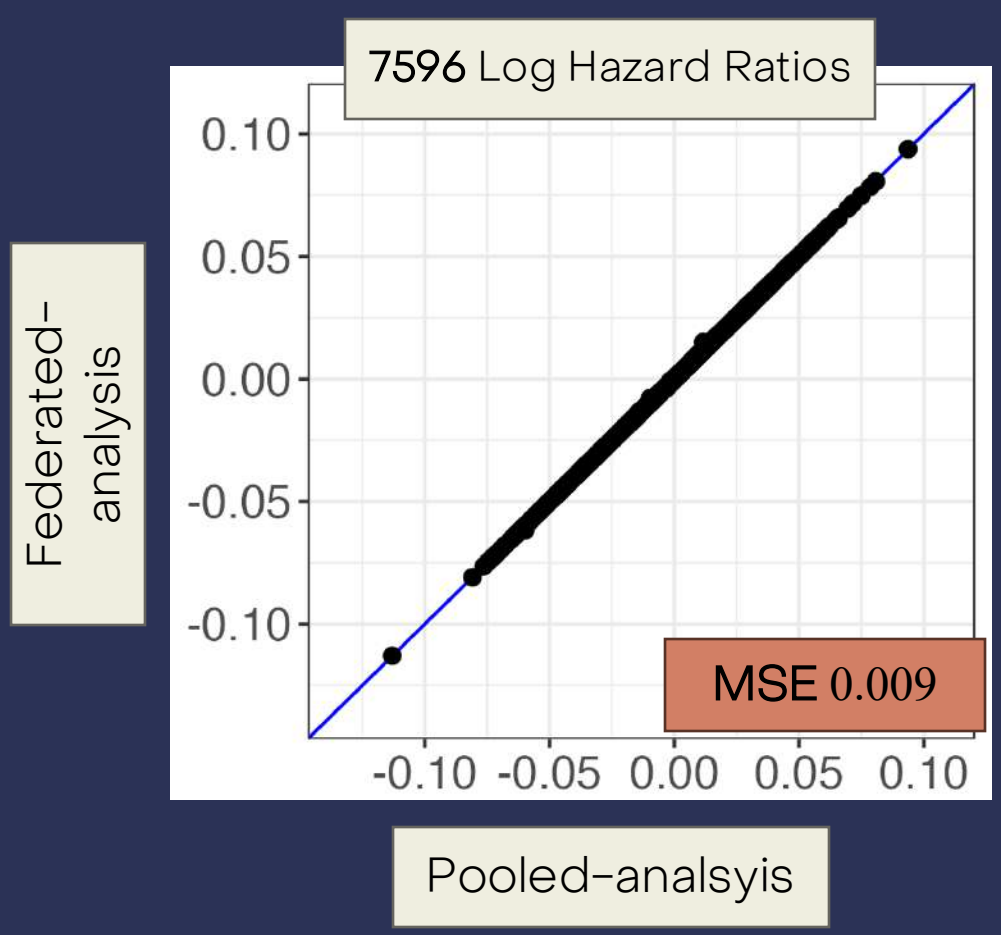
~10,000 observations

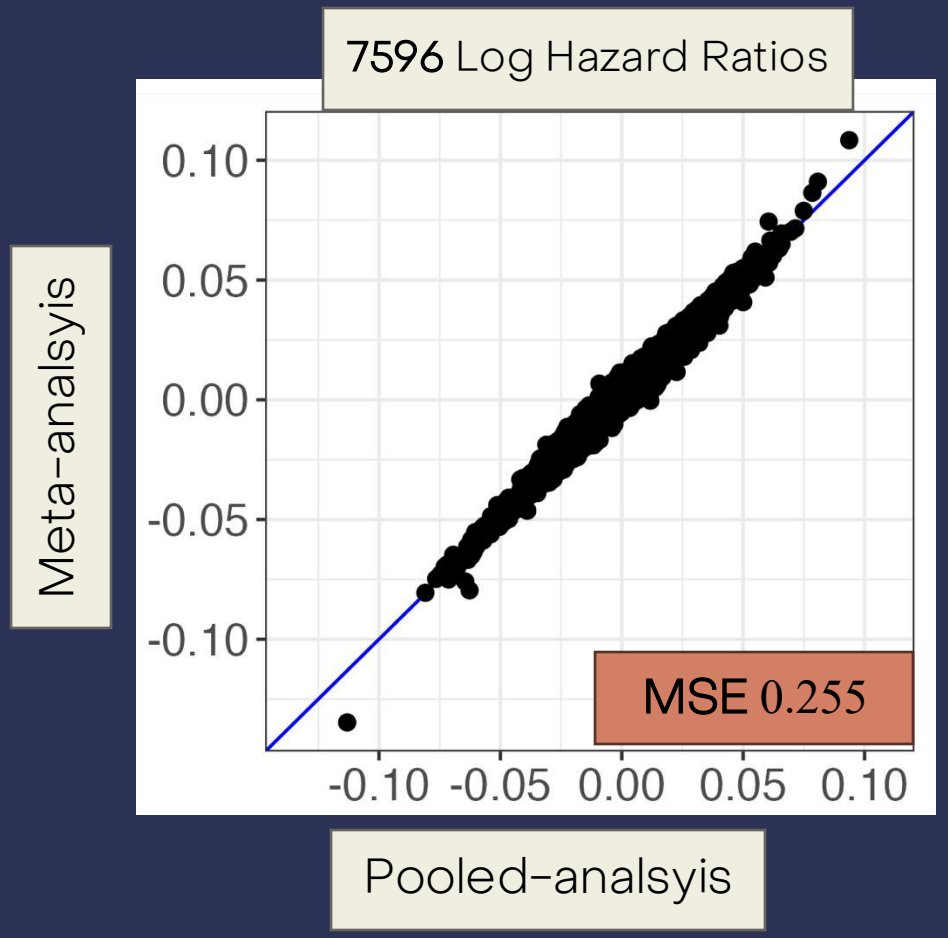
CoxPH model

Cancer ~ protein + Adjustment factors

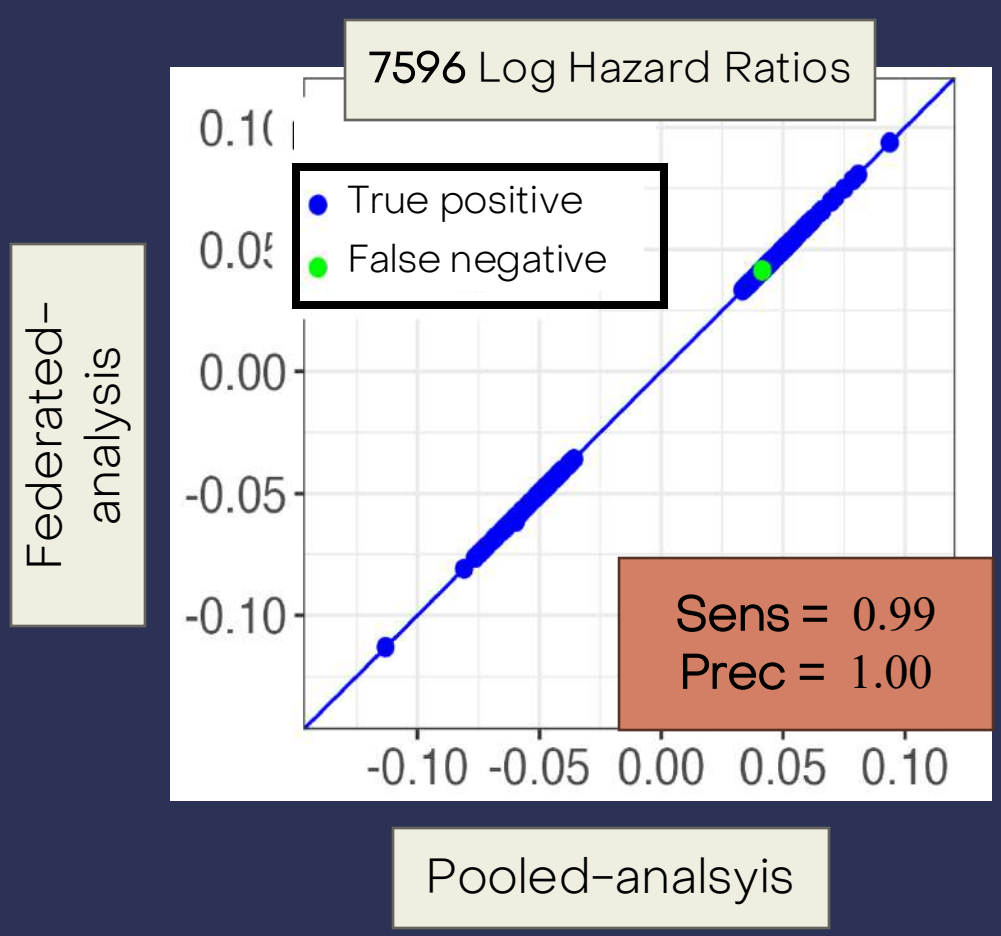
age, sex, etc

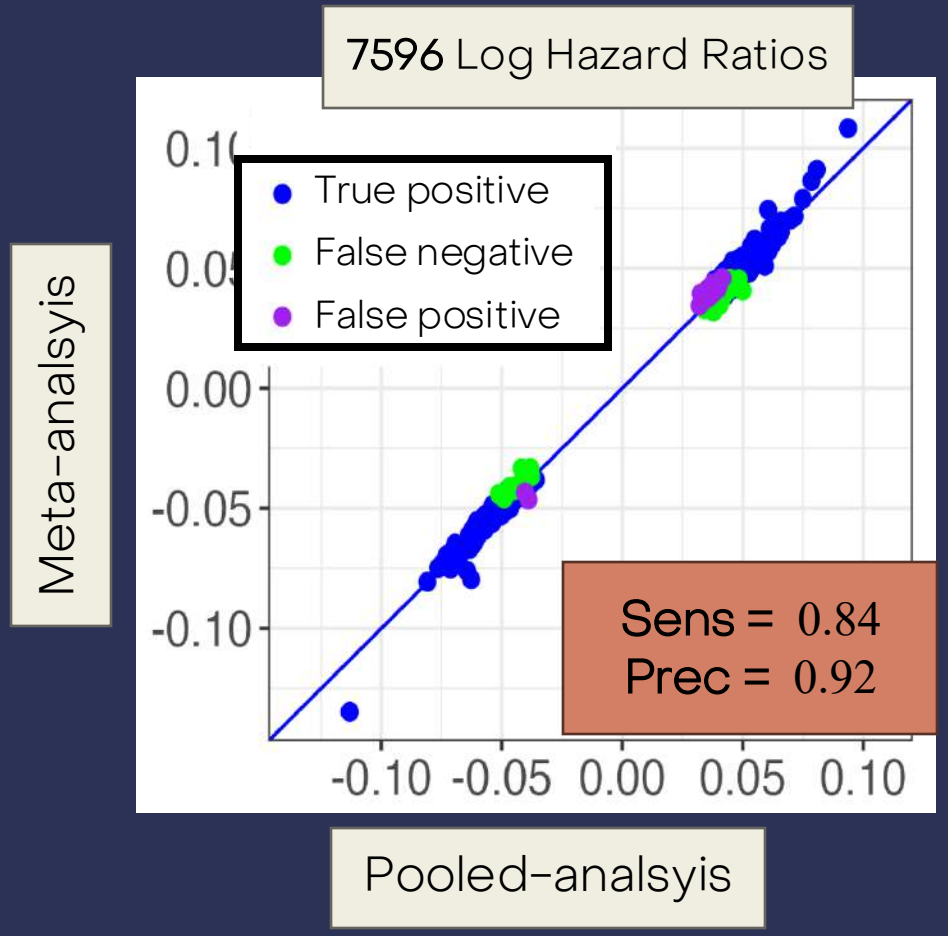
~7500 Hazard Ratios


7596 Log Hazard Ratios
Federated-analysis
0.10
0.05
0.00
-0.05
-0.10
MSE 0.009
-0.10 -0.05 0.00 0.05 0.10
Pooled-analsyis


7596 Log Hazard Ratios
Meta-analsyis
0.10
0.05
0.00
-0.05
-0.10
MSE 0.255
-0.10 -0.05 0.00 0.05 0.10
Pooled-analsyis

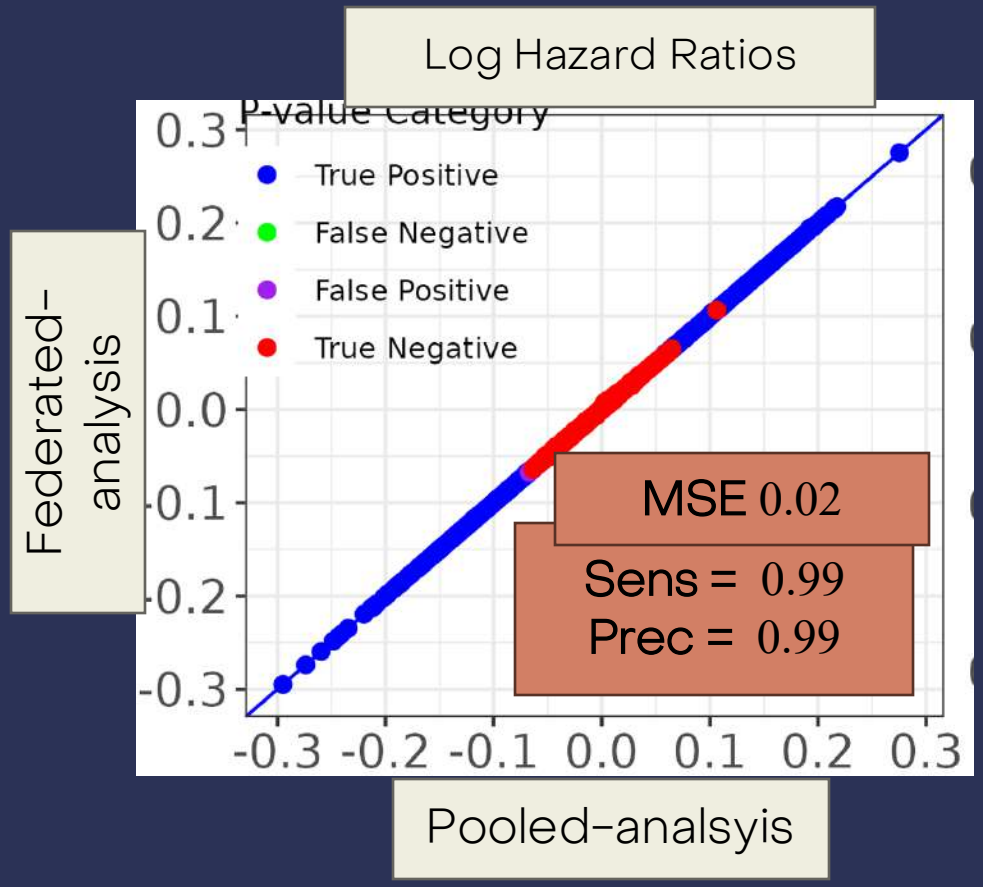
The Benjamini-Hochberg (FDR) adjusted P-Values

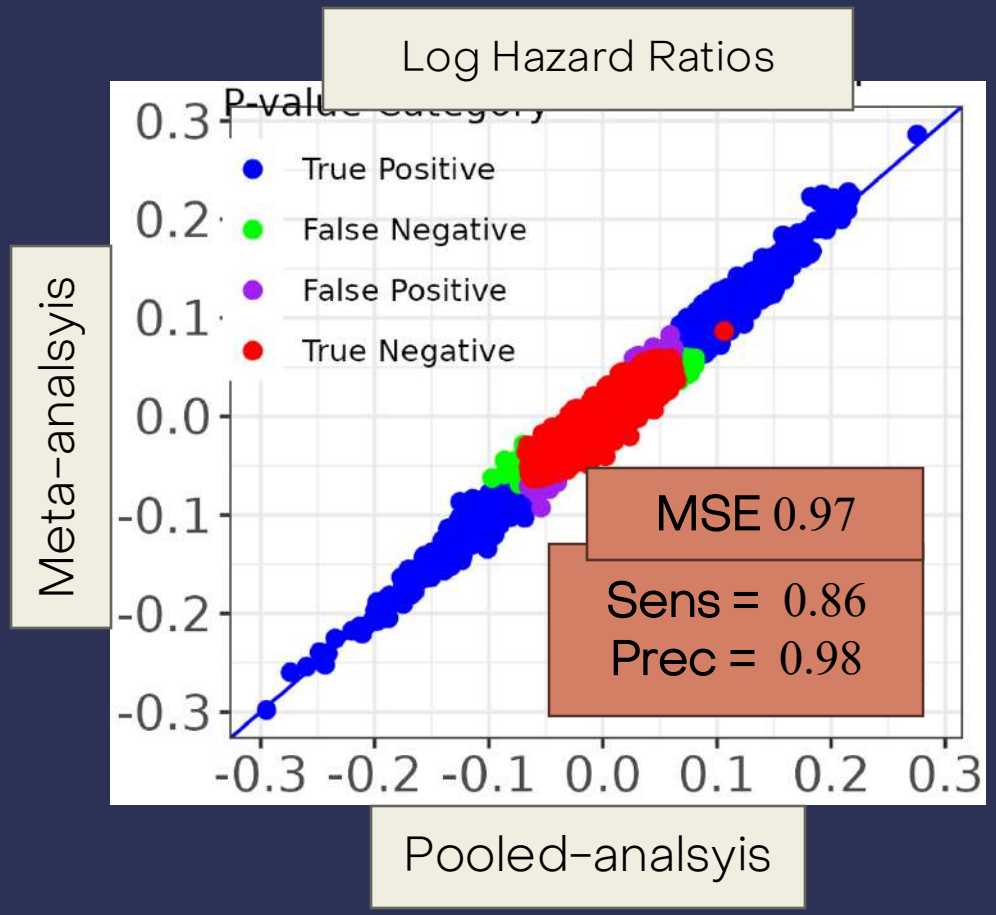




CoxPH models

Case-cohort design weights [under development]


Log Hazard Ratios
Federated-analysis
P-value Category
True Positive
False Negative
False Positive
True Negative
MSE 0.02
Sens = 0.99
Prec = 0.99
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.3 -0.2 -0.1 0.0 0.1 0.2 0.3
Pooled-analsyis


Log Hazard Ratios
Meta-analsyis
True Positive
False Negative
False Positive
True Negative
MSE 0.97
Sens = 0.86
Prec = 0.98
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.3 -0.2 -0.1 0.0 0.1 0.2 0.3
Pooled-analsyis

Feedback

Conclusion

Federated analysis:

- Privacy preserving
- Close results to pooled analysis
- Implementation is feasible

PDA:

Pros:

- Easy to setup and use.
- Open to development.

Cons:

- Limited list of methods.
- Mostly tested on synthetic data

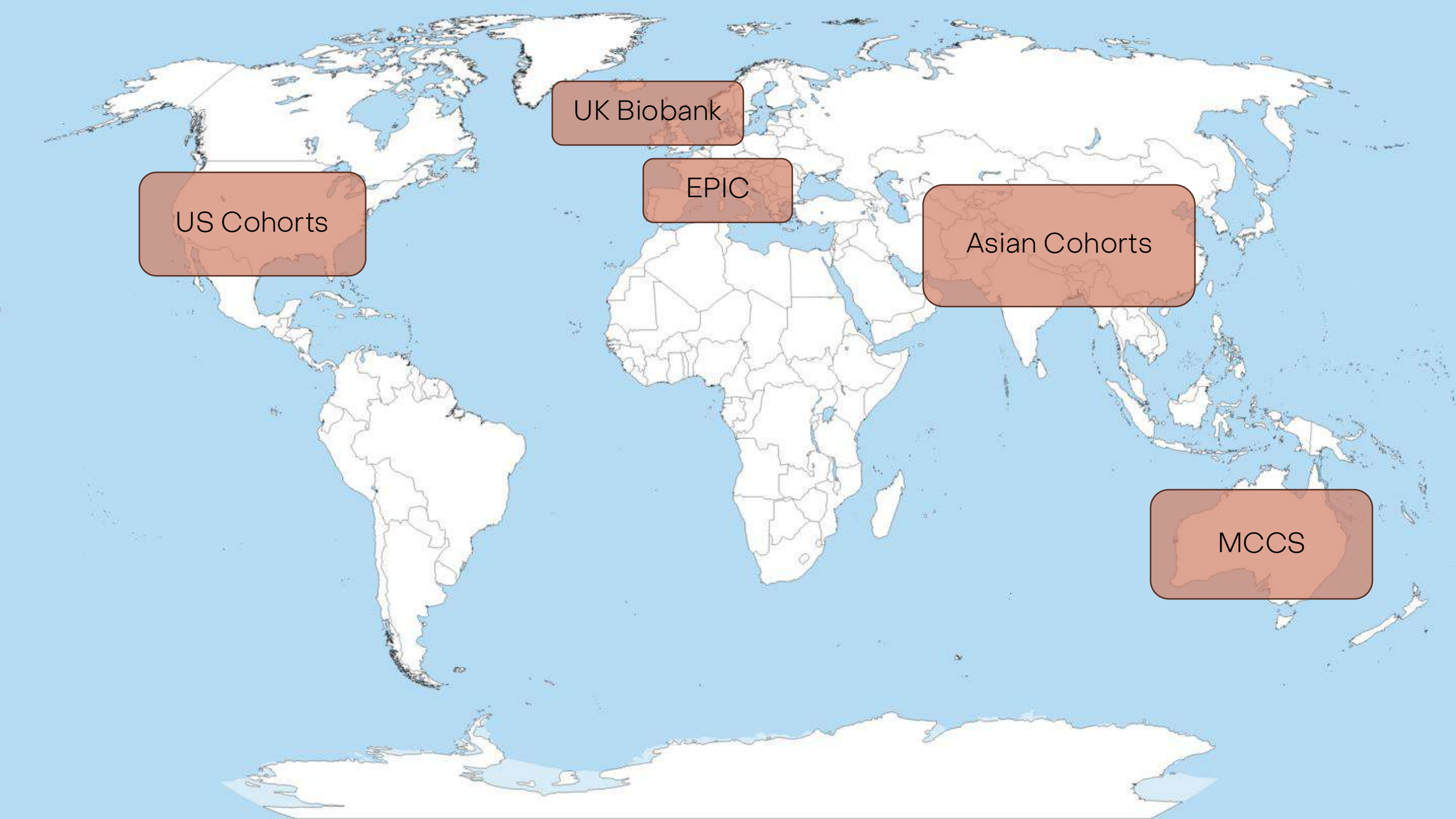
DataSHIELD:

- Automated.
- Adds complexity

International Agency for Research on Cancer

World Health Organization

Bonuses


UK Biobank
EPIC
US Cohorts
Asian Cohorts
MCCS

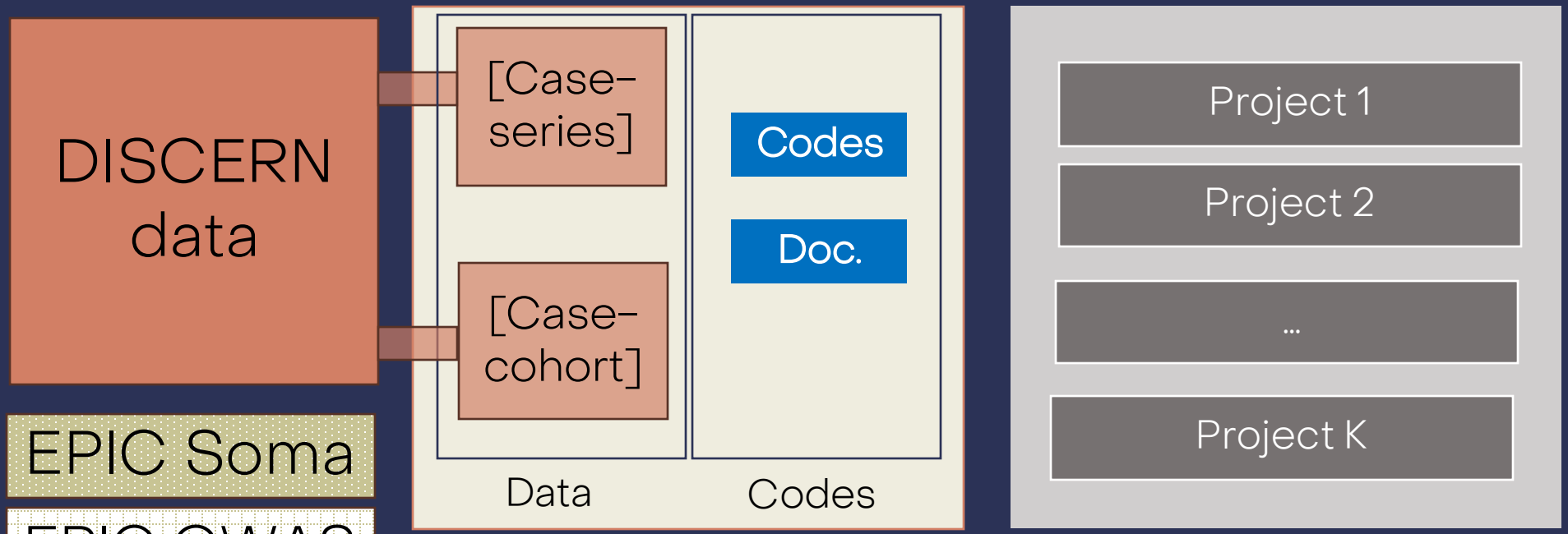
Next steps

- Debug and finalize variance estimation for case-cohort design using the PDA package.

- Investigate features and application of Data-Shield

- Publishing of a public website for all developed tools on a collaborative platform

DISCERN data and codes on the SIT platform



EPIC Soma

EPIC GWAS

Public: project space

Management

Shared

Subprojects

Data Managers

All registered users

Registered users (per project)

Consortium Folder

	Management	Private	Shared
Consortium PI	Read	Read	Read
Consortium Data Manager	Full	Full	Full
Consortium Private Member	None	Read	Read
Consortium Member	None	None	Read

SubProject Folders

	Sources	Files	Work
Consortium Data Manager	Full	None	None
Subproject PI	Read	Read	Full
Subproject Data Manager	Read	Full	Full
Subproject Member	Read	Read	Full