Implementing disclosure controls in DataSHIELD demonstrated by the dsSurvival package

DAGStat Conference 2022

Sofack Ghislain N.

Institute of Medical Biometry and Statistics (IMBI), University Medical Center Freiburg

27.03.2022

Survival analysis

Rationale

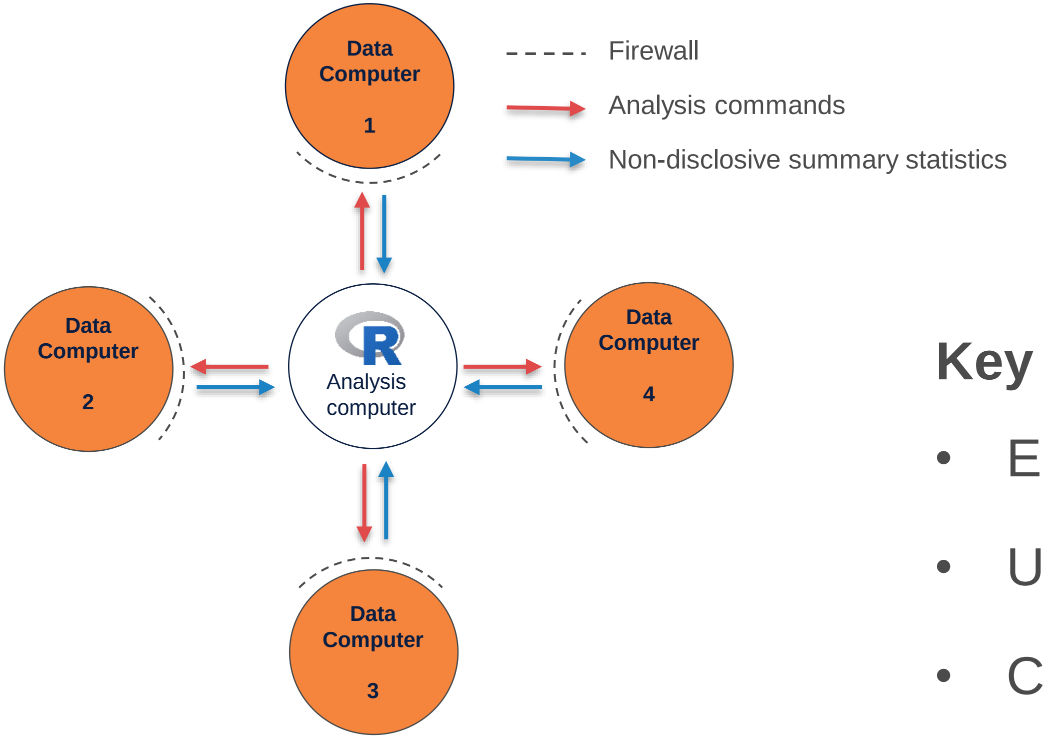
- Survival analysis is widely used in medical sciences to analyze the expected duration of time until some event of interest occurs
- The most frequently used model is the cox proportional hazard model (Cox, 1972)
- Performing meta-analysis of survival models requires large amount of data from different sites
  - General Data Protection Regulation
  - Physical size of data

Alternative: DataSHIELD

The DataSHIELD approach

Take “analysis to data” ….. not “data to analysis”

Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual-levEL Databases



Key principles

- Enables federated analysis
- Uses client – server architecture
- Controls disclosure risks

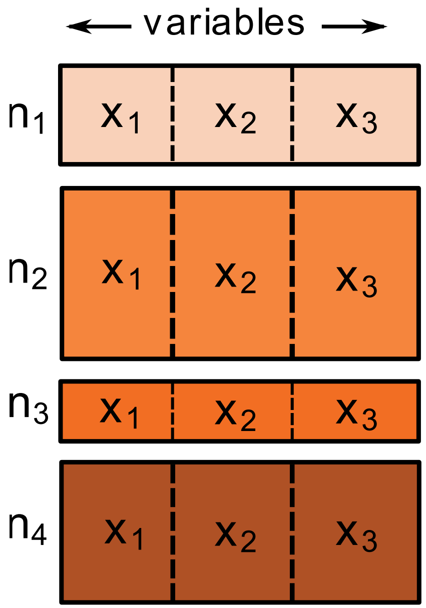
Gaye, Amadou, et al. "DataSHIELD: taking the analysis to the data, not the data to the analysis." International journal of epidemiology 43.6 (2014): 1929-1944.

https://www.datashield.org/

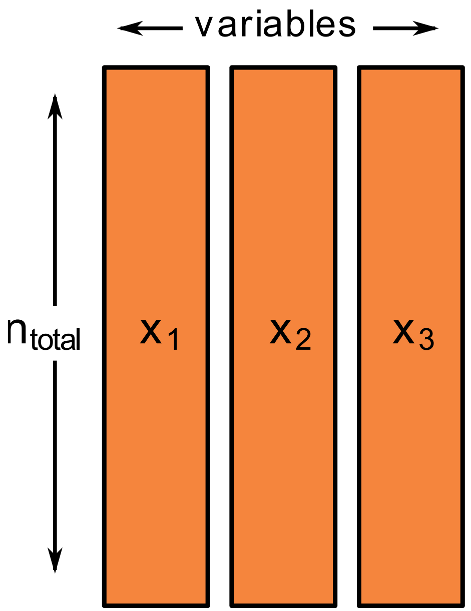
The DataSHIELD approach

Two classes of multi-score analysis

- **Horizontal partitioning**
  - meta-analysis setting



- Vertical partitioning
  - record linkage setting



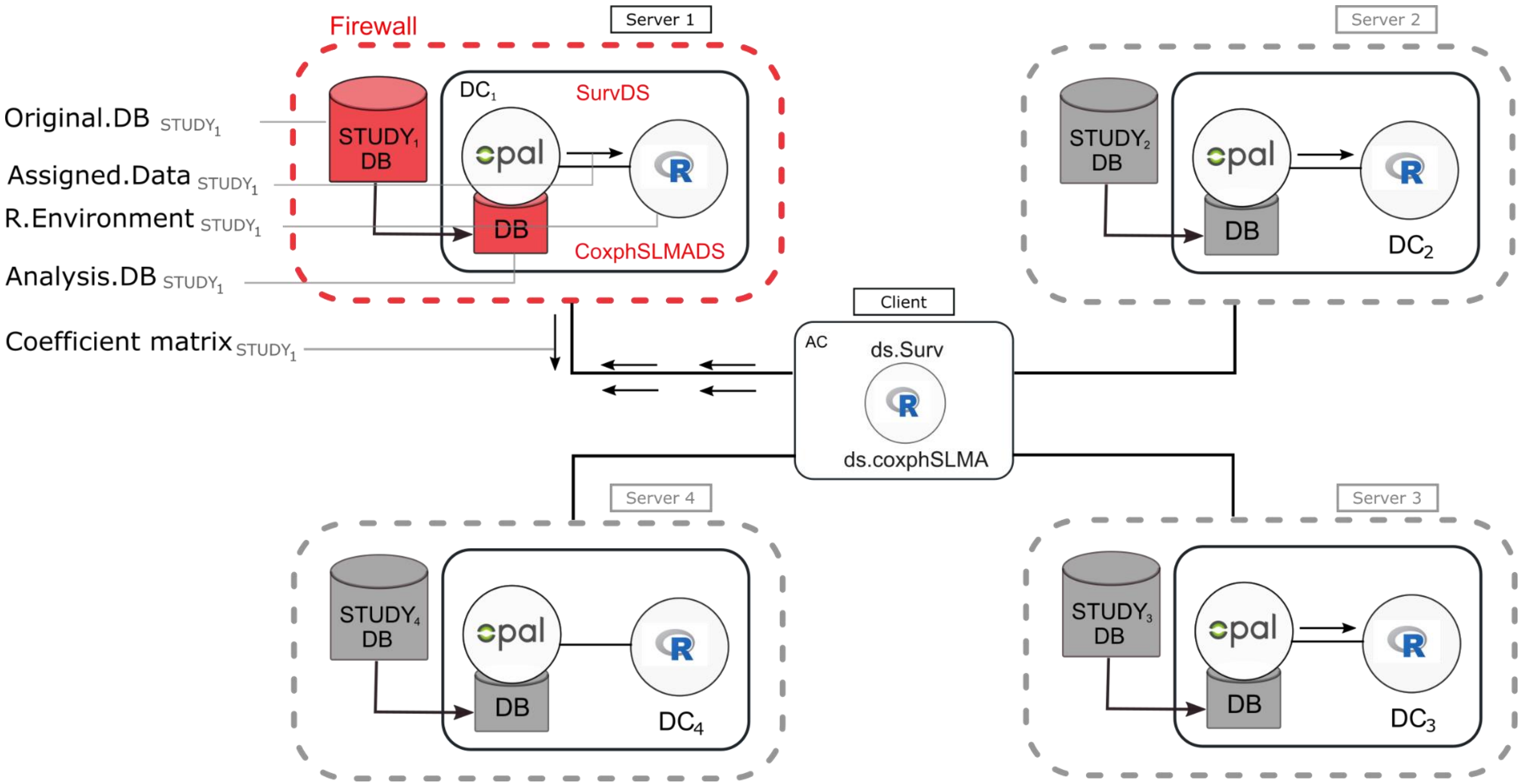
dsSurvival

Privacy preserving fitting of Cox models

- Allow Cox models to be fitted at each study, and then meta analyse the results
- Implementation is restricted to being study-level meta-analysis (SLMA) rather than full likelihood
- Server-side package: dsSurvival
  - SurvDS(…)
  - coxphSLMADS(…)
- Client-side package: dsSurvivalClient
  - ds.Surv(…) → assign function
  - ds.coxphSLMA(…) → aggregate function

dsSurvival Framework

Privacy preserving fitting of Cox models



AC: Analysis computer
DC: Data computer
DB: Database

Disclosure risks

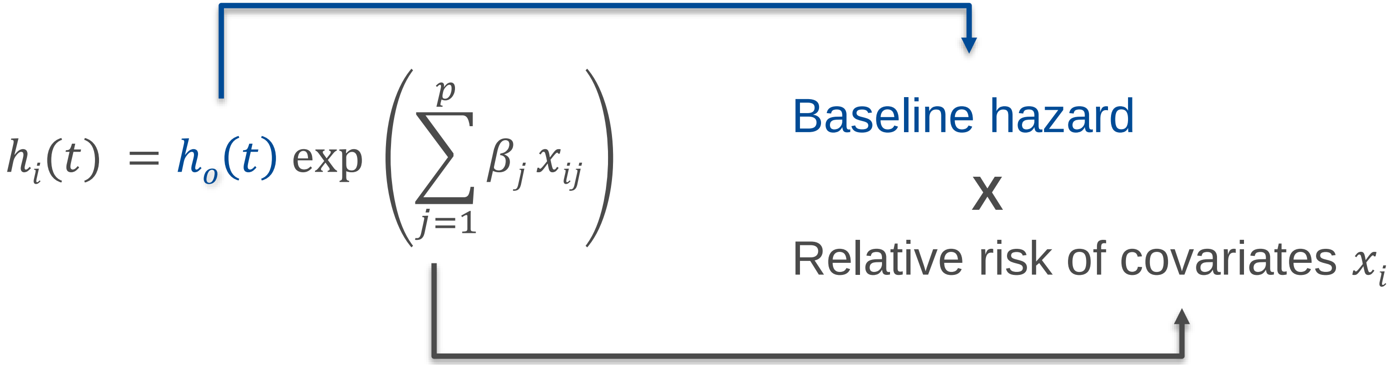
Survival analysis

- Controlling the risk that the data analyst can deliberately infer to the identity or to one of the key variables being analyzed.
- The results of a survival analysis are likely to be disclosive if:
  - Reveal identifying information, or exact values of variables, including dates, diagnoses, and comorbidities
  - Reveal status of observations

O'Keefe, Christine M., et al. "Confidentialising survival analysis output in a remote data access system." Journal of Privacy and Confidentiality 4.1 (2012).

Disclosure risks

Cox proportional hazard models



$$h_i(t) \ = h_o(t) \exp\left(\sum_{j=1}^{p} \beta_j \, x_{ij}\right)$$

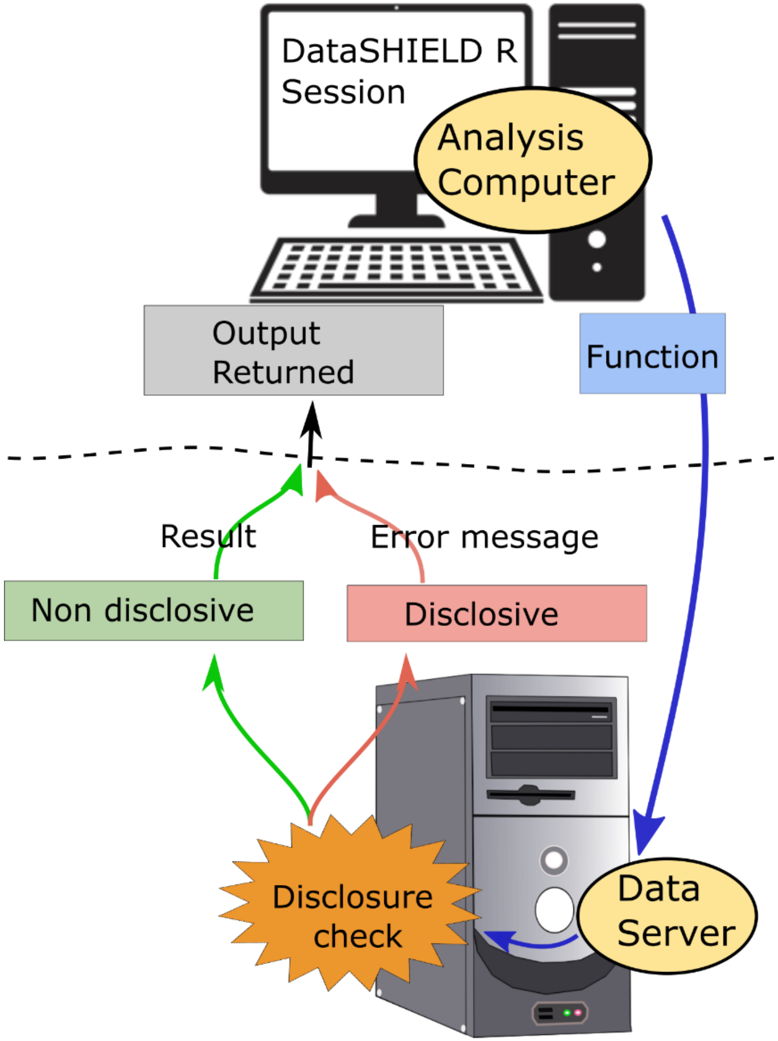
Baseline hazard

X

Relative risk of covariates x_i

- Interested in the coefficient estimates β rather than the baseline hazard $h_o(t)$
- Do not release the values of the covariates x_{ij} for each participant
- Do not reveal the hazard function $h_i(t)$ (survival objects) for each participant

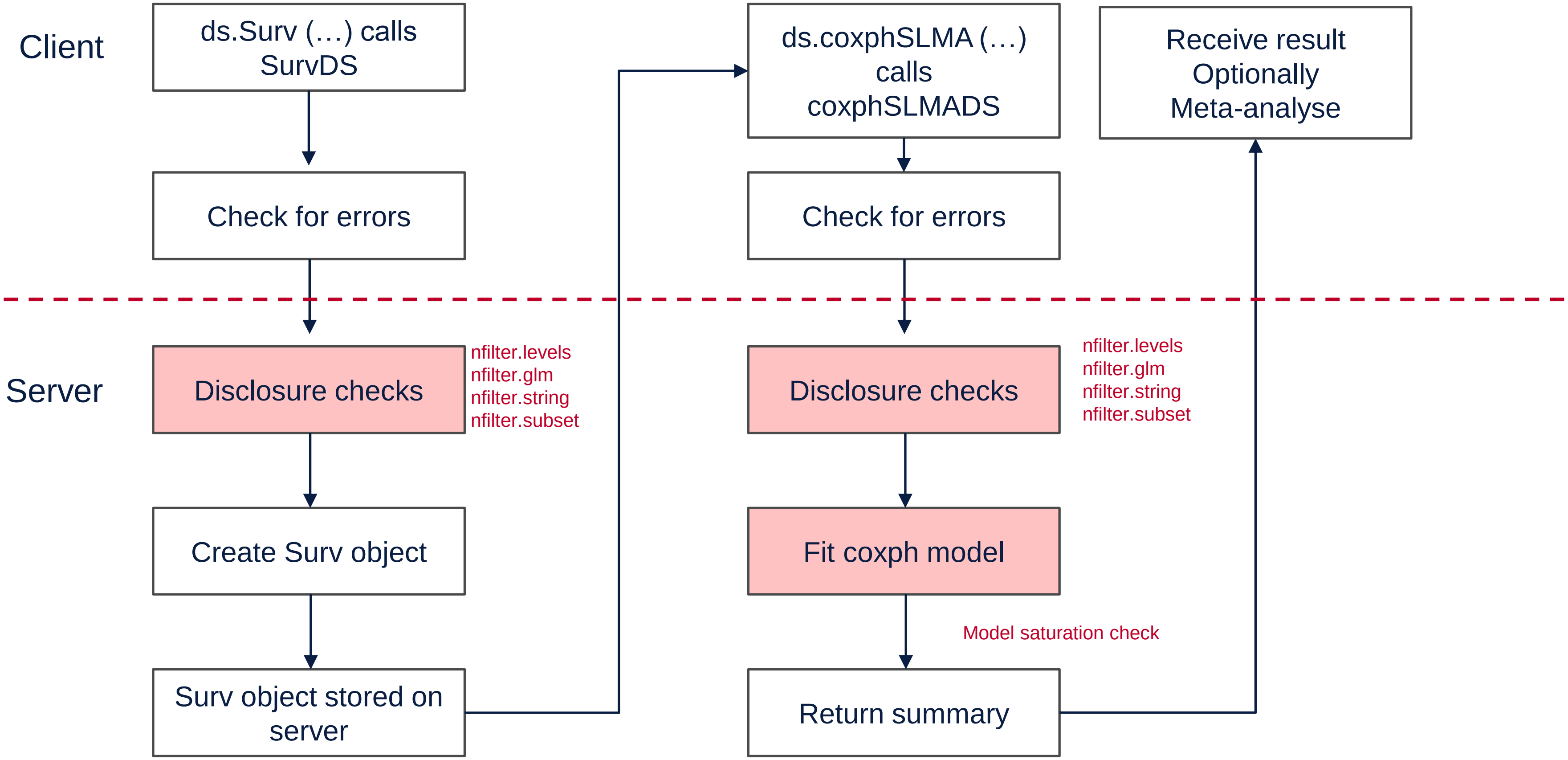
Disclosure control



Disclosure checks

- nfilter.levels
- nfilter.tab
- nfilter.glm
- nfilter.string
- nfilter.subset

Disclosure checks



- Number of parameters in Cox model as a proportion of the sample size
- Default : 20% of sample size
- Prevents model oversaturation

Output presentation

```
$study1

                     coef       exp(coef)   se(coef)   z        Pr(>|z|)
 D$age               0.00815    1.008191    0.001248   6.535    6.35e-11 ***
 D$bmi               0.00553    1.005551    0.030356   2.422    0.004245 **
 D$factor(sex)male   0.15224    1.164442    0.065621   0.215    0.000116 **

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1‘ ’ 1
```

```
$study2

                     coef       exp(coef)   se(coef)   z        Pr(>|z|)
 D$age               0.04067    1.04151     0.00416    9.776    < 2e-16  ***
 D$bmi               -0.62756   0.53389     0.11767    -5.333   9.66e-08 ***
 D$factor(sex)male   -0.66000   0.516850    0.099481   -6.634   3.26e-11 ***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1‘ ’ 1
```

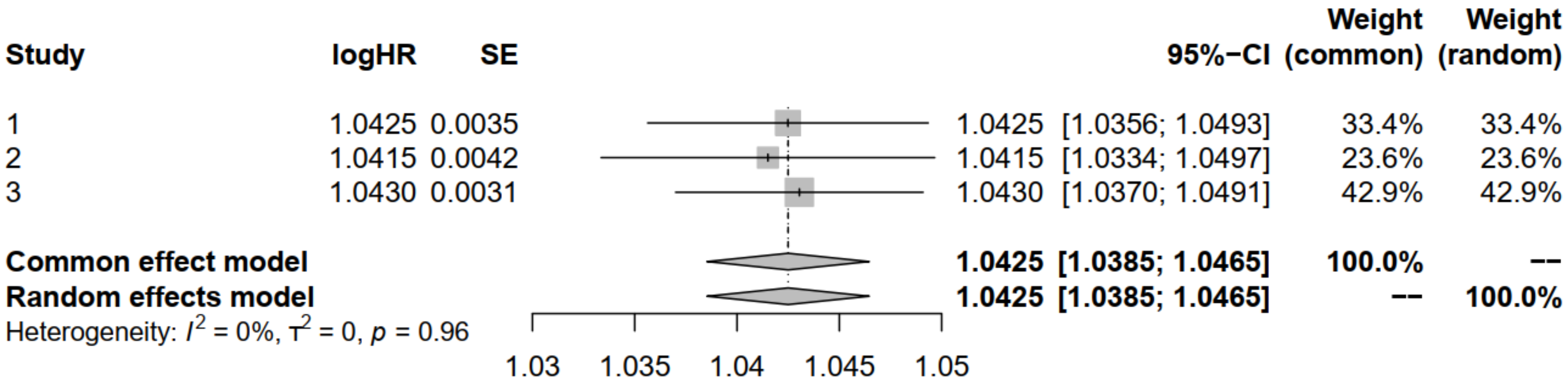
```
$study3

                     coef       exp(coef)   se(coef)   z        Pr(>|z|)
 D$age               0.042145   1.043045    0.003086   13.655   < 2e-16  ***
 D$bmi               0.006522   1.005551    0.03359    1.452    0.424513 ***
 D$factor(sex)male   -0.599238  0.549230    0.084305   -7.108   1.18e-12 ***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1‘ ’ 1
```

Metafor R package

- Meta-analysis of the hazard ratios
- Forest plots of estimates from RE model



Study	logHR	SE		95%-CI	Weight (common)	Weight (random)
1	1.0425	0.0035	1.0425	[1.0356; 1.0493]	33.4%	33.4%
2	1.0415	0.0042	1.0415	[1.0334; 1.0497]	23.6%	23.6%
3	1.0430	0.0031	1.0430	[1.0370; 1.0491]	42.9%	42.9%
Common effect model			**1.0425**	**[1.0385; 1.0465]**	**100.0%**	**--**
Random effects model			**1.0425**	**[1.0385; 1.0465]**	**--**	**100.0%**

Heterogeneity: $I^2 = 0\%$, $\tau^2 = 0$, $p = 0.96$

http://www.metafor-project.org

Summary

- DataSHIELD enables federated analysis and tailored disclosure controls
- dsSurvival is a DataSHIELD package for privacy preserving meta-analysis of survival data distributed across different sites
- A tutorial in bookdown format with code, diagnostics, plots and synthetic data is available here:
- https://neelsoumya.github.io/dsSurvivalbookdown/
- All code is available from the following repositories:
- https://github.com/neelsoumya/dsSurvivalClient/
- https://github.com/neelsoumya/dsSurvival/

Thank you

- Daniela Zöller
- Soumya Banerjee
- Thodoris Papakonstantinou
- Tom R.P. Bishop
- Paul Burton
- Demetris Avraam