Creating high utility synthetic data (with synthpop)

Preserving relationships between variables

- Preserving solely univariate distributions is not always sufficient
- Statistics Netherlands has awesome data that you can use to answer many research questions
- But… these data are often not (easily) accessible

- Synthetic data can provide a solution, but only if relationships between variables are preserved

Synthpop

- Developed by Beata Nowok, Gillian Raab and Chris Dibben (2016)
- Default: classification and regression trees (**CART**)
- Preserving relationships by a sequential regression framework

- $X_1^{\{syn\}} \sim Sample\left(X_1^{\{obs\}}\right)$
- $X_2^{\{syn\}} \sim CART\left(X_2^{\{obs\}} \middle| X_1^{\{syn\}}\right)$
- $X_K^{\{syn\}} \sim CART\left(X_K^{\{obs\}} \middle| X_1^{\{syn\}}, \ldots, X_{K-1}^{\{syn\}}\right)$

Privacy

- **Higher privacy risk:** CART reuses observed values

- How to quantify this remains an open question

Utility

- Higher utility by adding more information in the synthesis model

General utility

- $pMSE$: predict which observations are real, and which are synthetic

Specific utility

Enough talking, start doing

- Hands-on session: https://thomvolker.github.io/osf_synthetic/osf_synthetic_workshop.html

- R Studio or R Studio Cloud

- Own data or *Heart failure clinical records* data (uploading your own data to R Studio Cloud might not be the best idea)

- Any questions?