The Avatar method: computation of synthetic data and application in health

Olivia Rousseau, CR2TI, Nantes Université

Conflict of interest

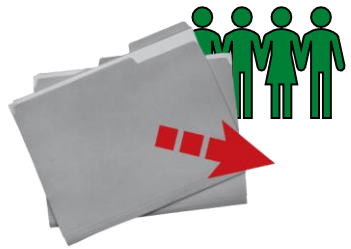
PA Gourraud is the founder of Methodomics(2008) and the co-founder of Big data Santé (2018). He consults for major pharmaceutical companies,and start-ups, all of which are handled through academic pipelines (AstraZeneca, Biogen, BostonScientific, Cook, Docaposte, Edimark, Ellipses, Elsevier, Methodomics, Merck, Mérieux, Octopize,Sanofi-Genzyme). PA Gourraud is a volunteer board member atAXA non-for-profit mutual insurancecompany (2021). He has no prescription activity with either drugs or devices

Type of data

Sensitive data

Pseudonymized data

Anonymized/Synthetic data

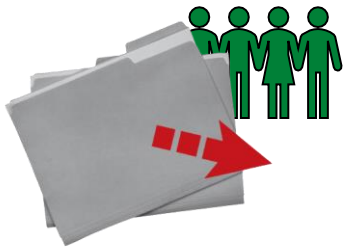


Type of data

Sensitive data

Pseudonymized data

Anonymized/Synthetic data

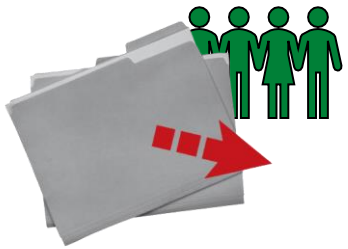


Type of data

Sensitive data

Pseudonymized data

Anonymized/Synthetic data



Major issue: Reidentification risk

- GDPR: General Data Protection Regularization

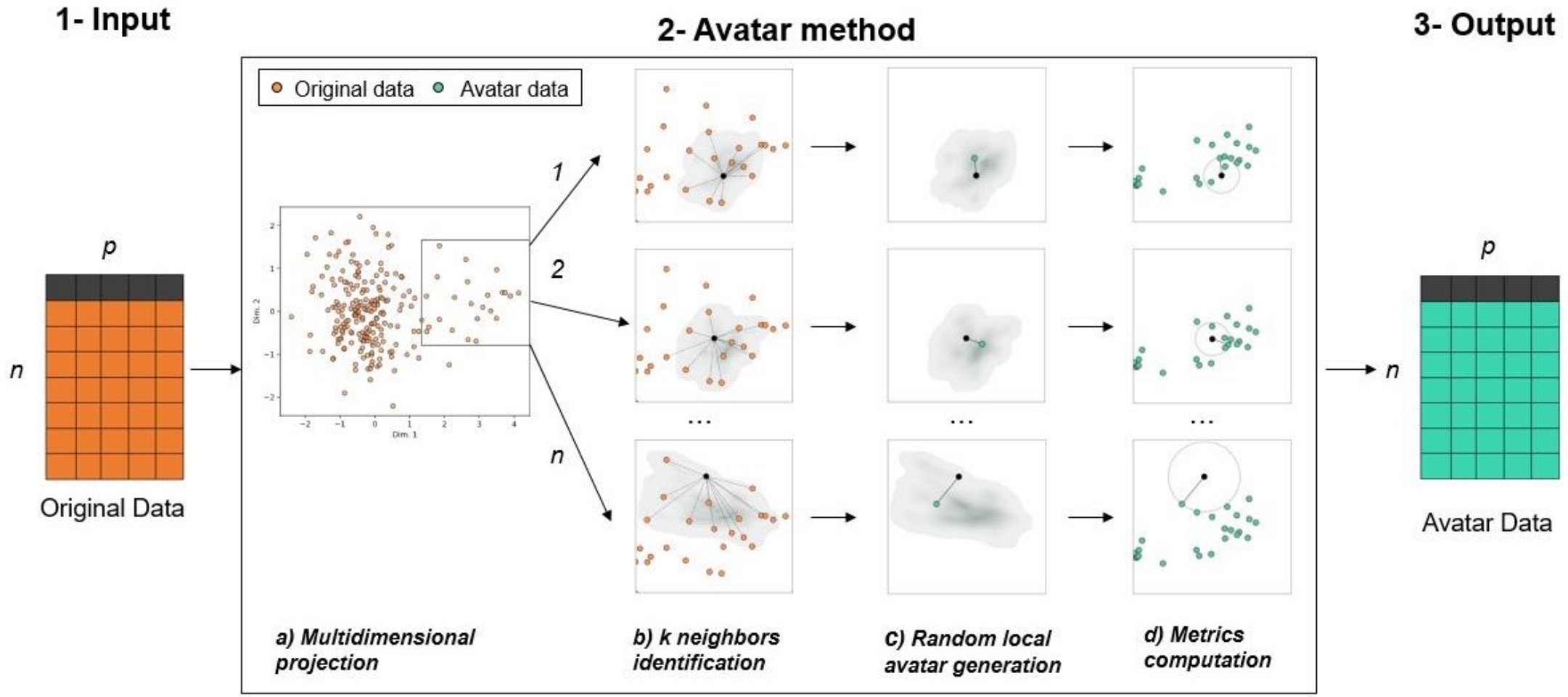
- Evaluation of the robustness of an anonymization process:
  - Singling out
  - Linkability
  - Inference

https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

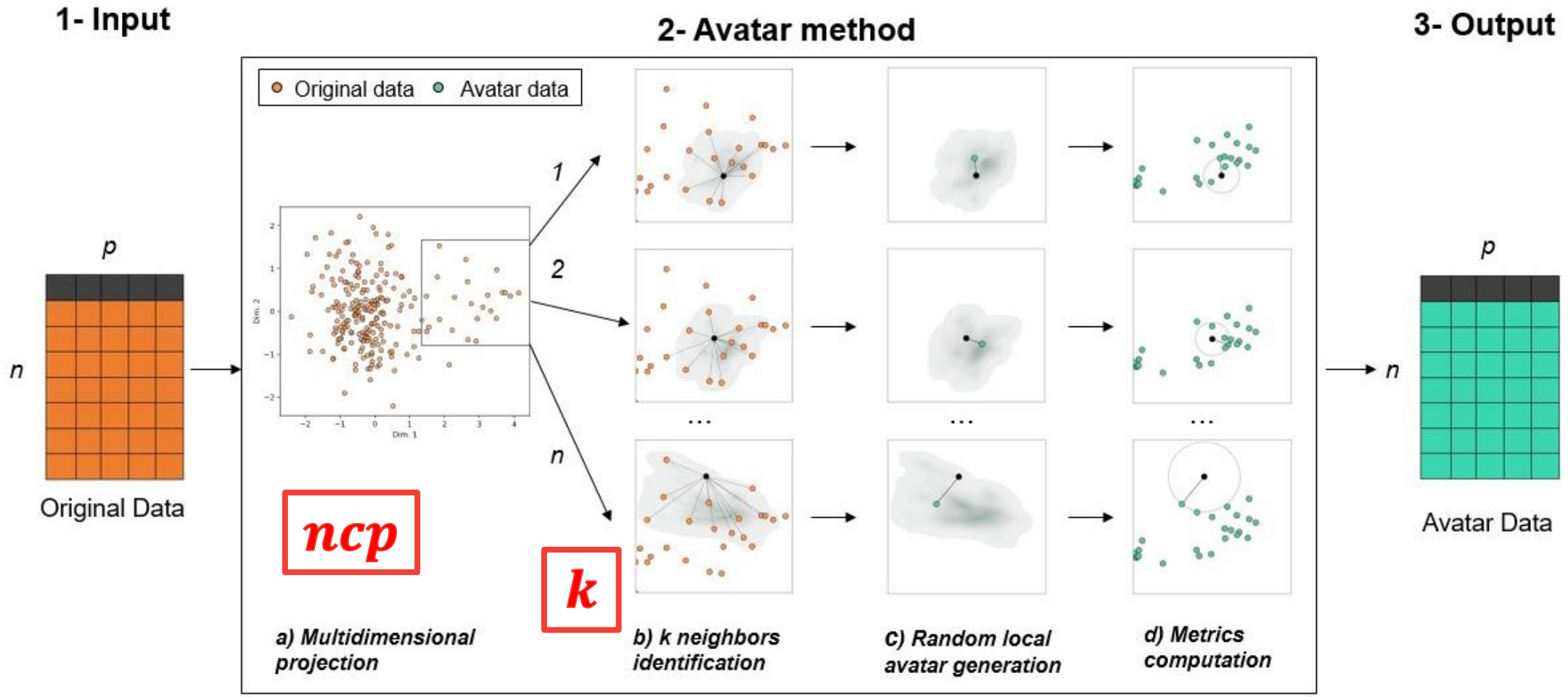
Avatar method

Patient-centric approach

Overview of the Avatar method

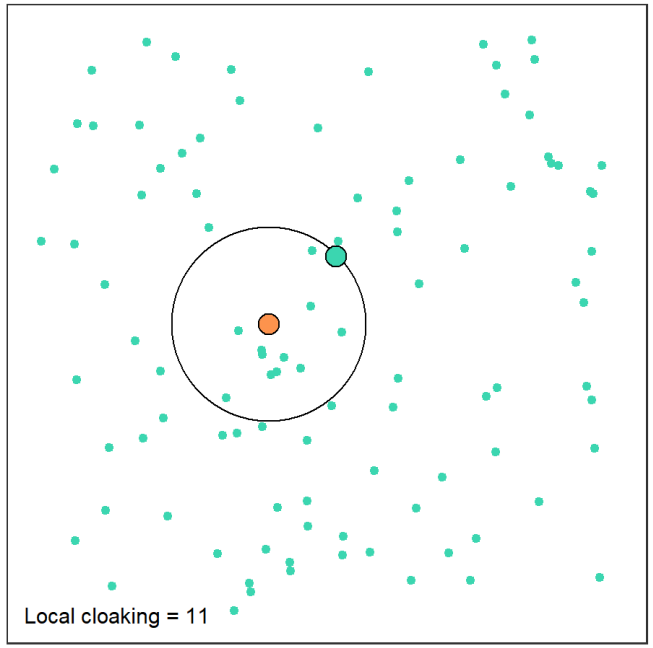


Overview of the Avatar method

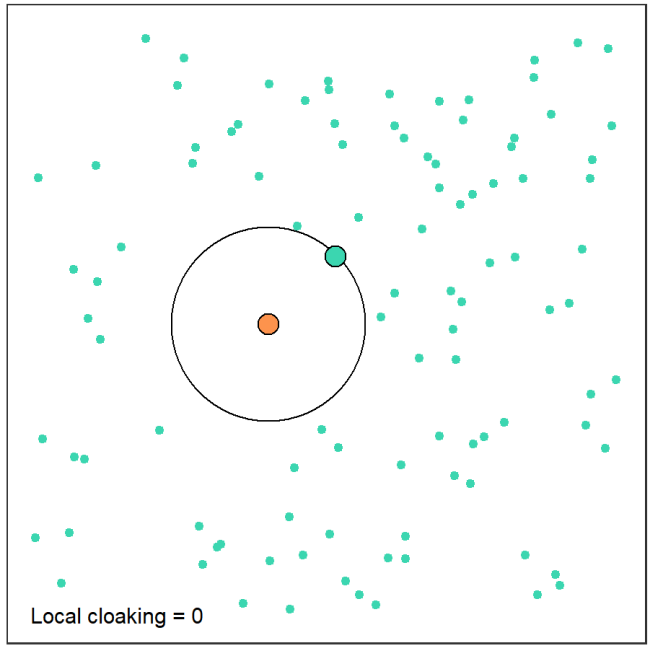


Security metrics

- Local cloaking



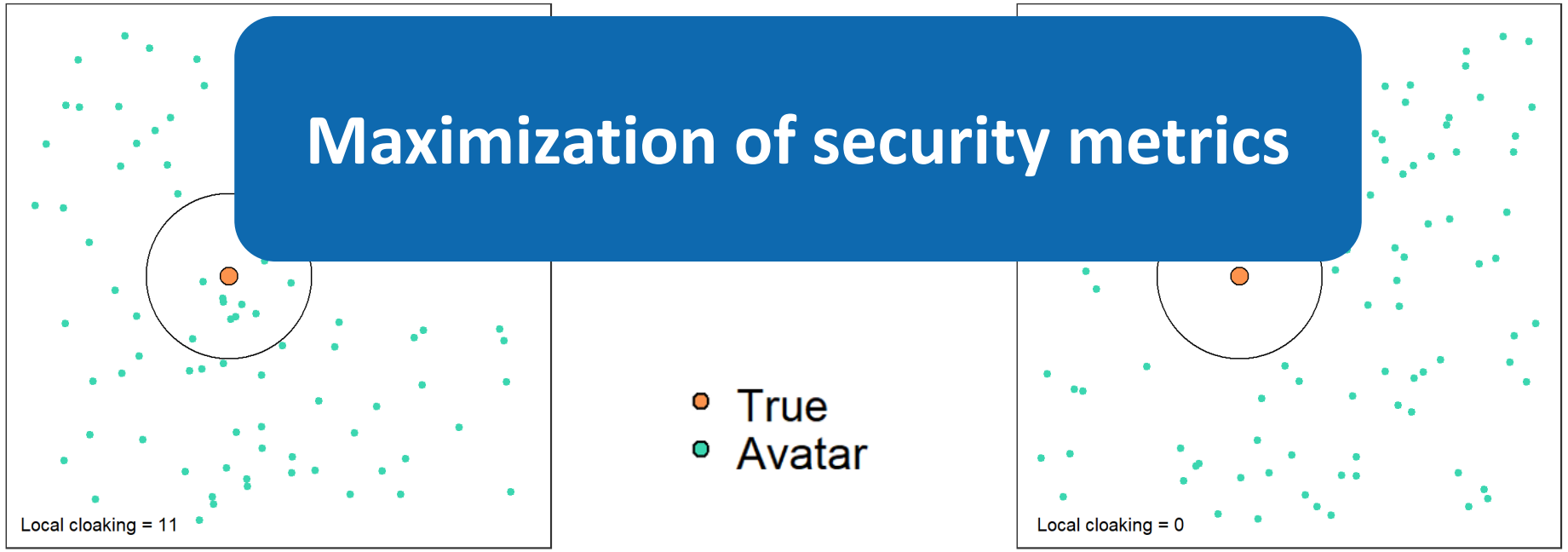
- Hidden rate



True
Avatar

Security metrics

- Local cloaking
- Hidden rate

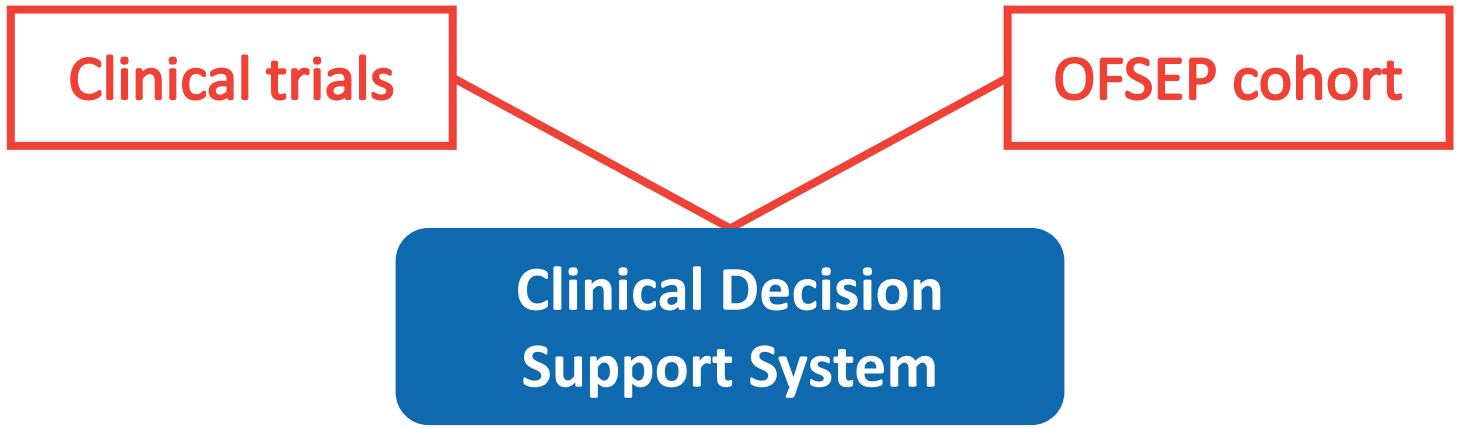


Application in health

PRIMUS project

PRIMUS project

- Multiple sclerosis: autoimmune disease



- **Objective:** Facilitate treatment choice

Data

- Clinical trial REFLEX :
  - Interferon beta-1a treatment doses study
- 518 patients
- 14 variables: 4 categorical and 10 continuous

- Treatments:
  - 172 patients: Placebo (Arm 0)
  - 171 patients: Treatment 3 time per week (Arm 1)
  - 175 patients: Treatment 1 time per week (Arm 2)

https://doi.org/10.1016/S1474-4422(11)70262-9

Data

- Clinical trial REFLEX :
  - Interferon beta-1a treatment doses study
- 518 pati
- 14 vari
- Treatme

Compare true and avatarized data from published results

  - 172 patients: Placebo (Arm 0)
  - 171 patients: Treatment 3 time per week (Arm 1)
  - 175 patients: Treatment 1 time per week (Arm 2)

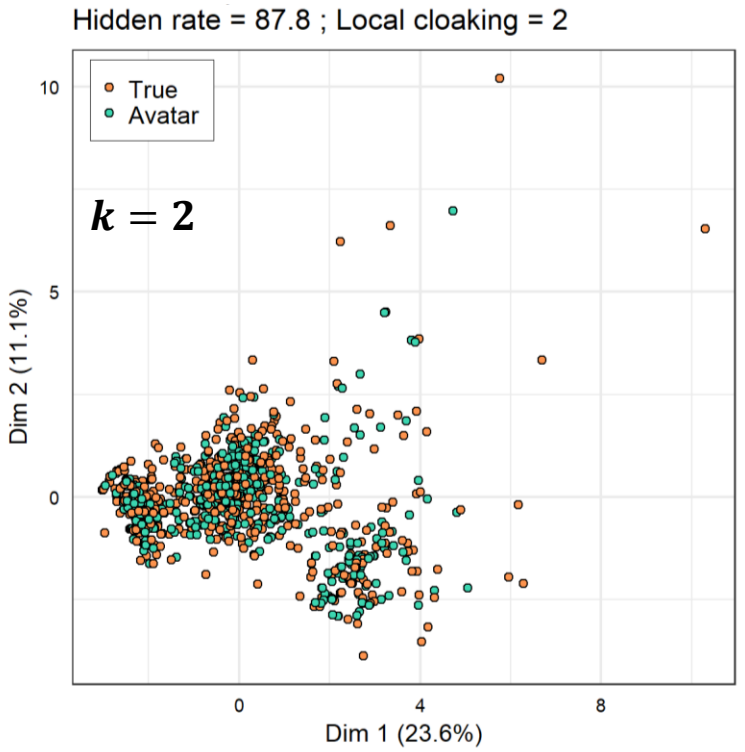
https://doi.org/10.1016/S1474-4422(11)70262-9

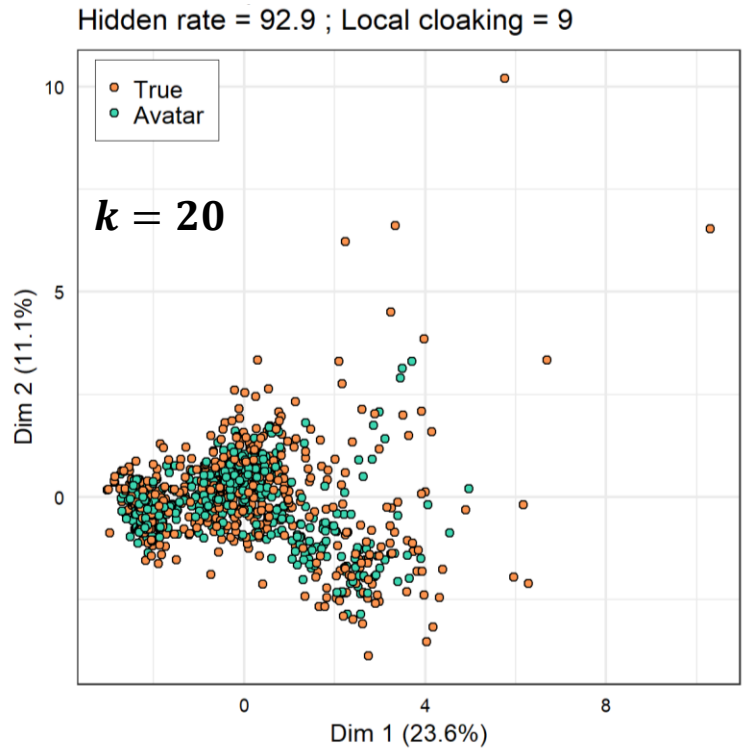
Avatar method – Parameters tests

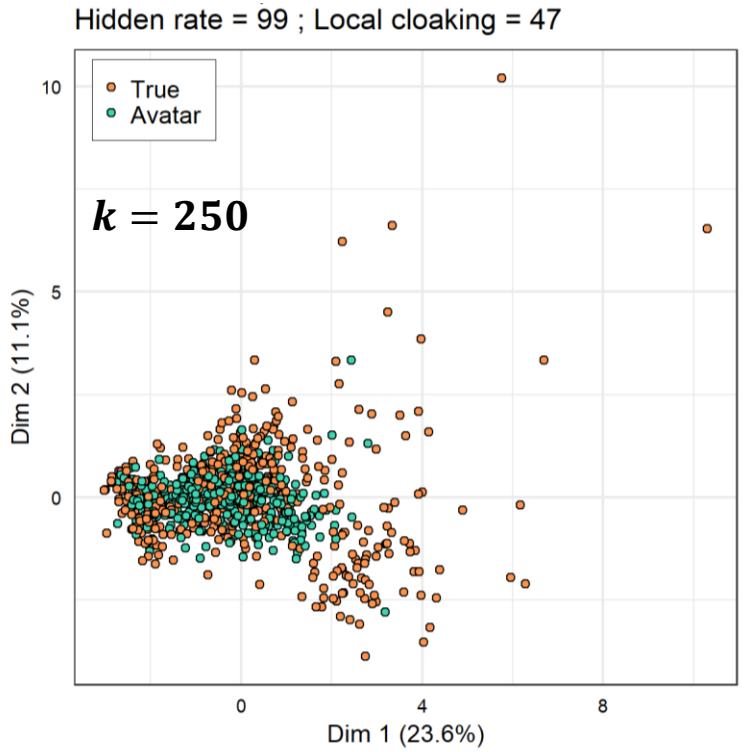
k

- Number of neighbors: $k \in \{2, 5, 10, 20, 50, 100, 250\}$

$ncp = 8$
$\%var = 80$





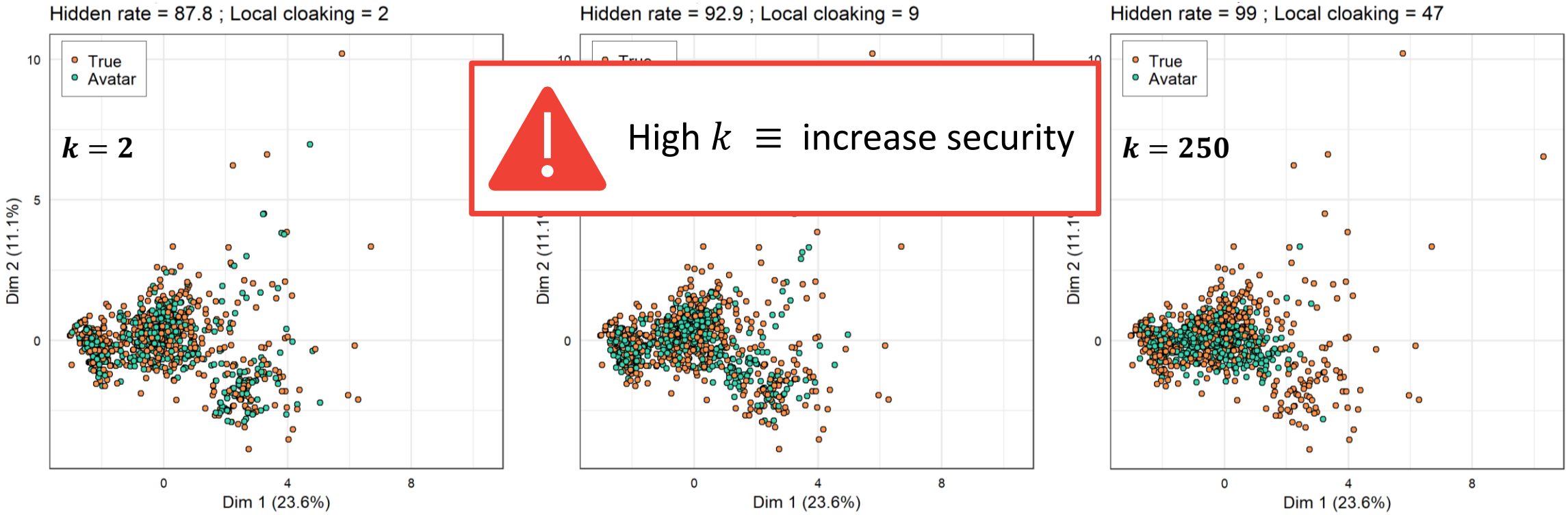


Avatar method – Parameters tests

k

- Number of neighbors: $k \in \{2, 5, 10, 20, 50, 100, 250\}$

$ncp = 8$
$\%var = 80$



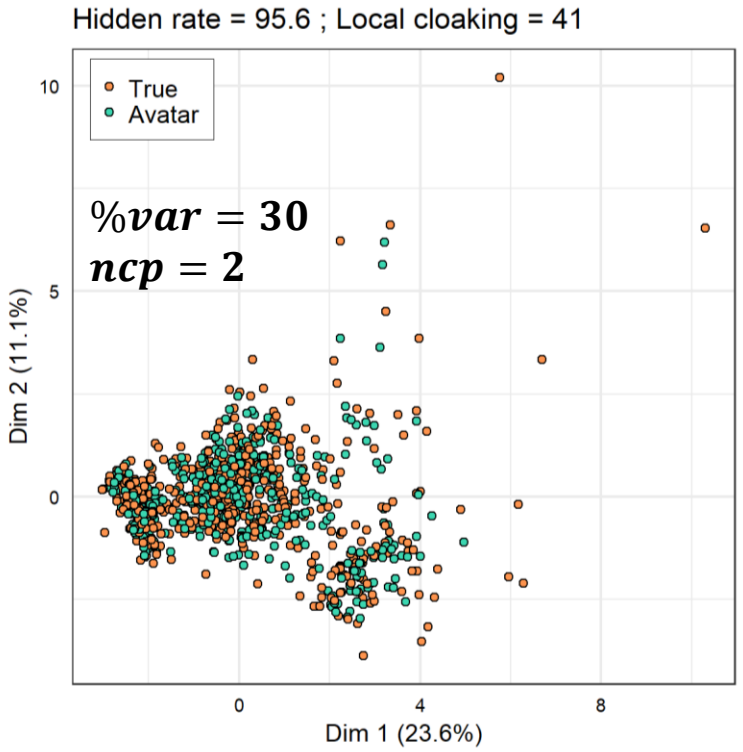
High k ≡ increase security

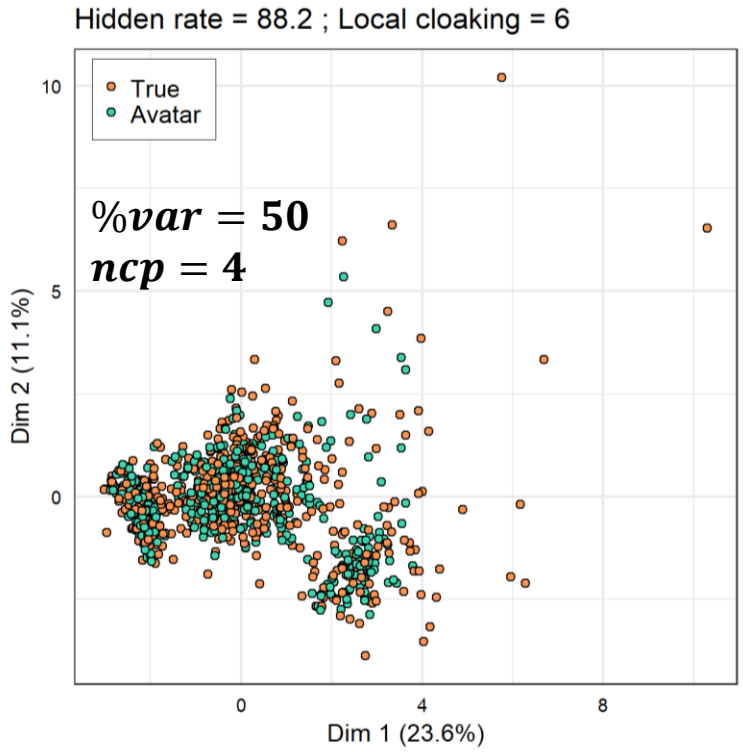
Avatar method – Parameters tests

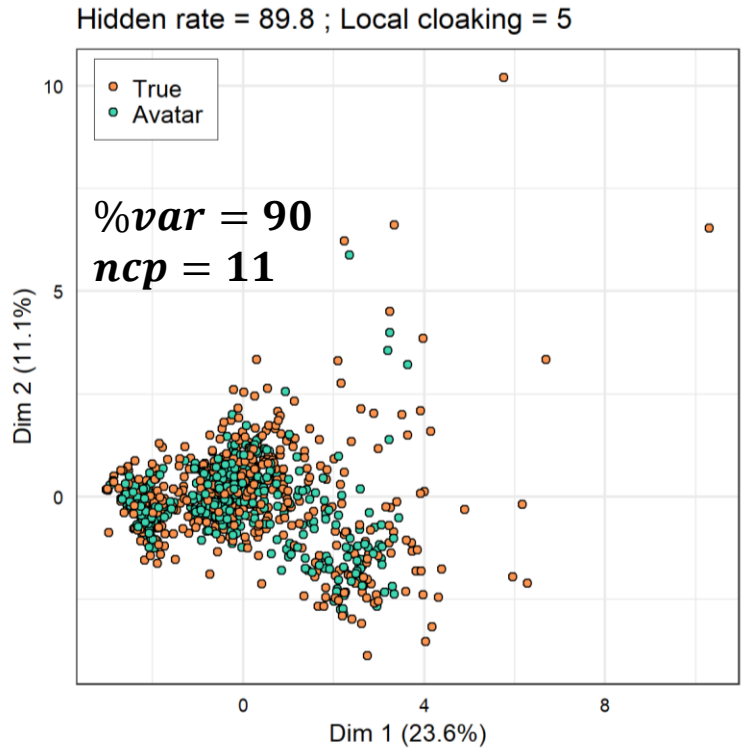
ncp

- % of explained variance: $\%var \in \{30, 50, 80, 90, 100\}$

$k = 10$





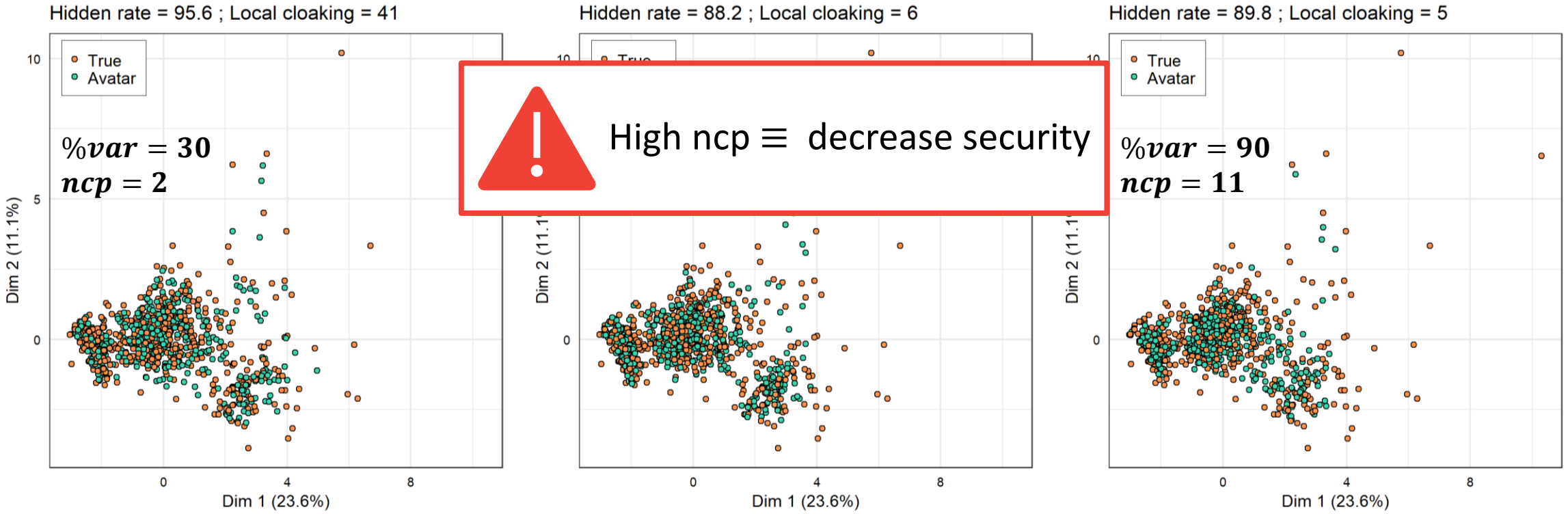


Avatar method – Parameters tests

ncp

- % of explained variance: $\%var \in \{30, 50, 80, 90, 100\}$

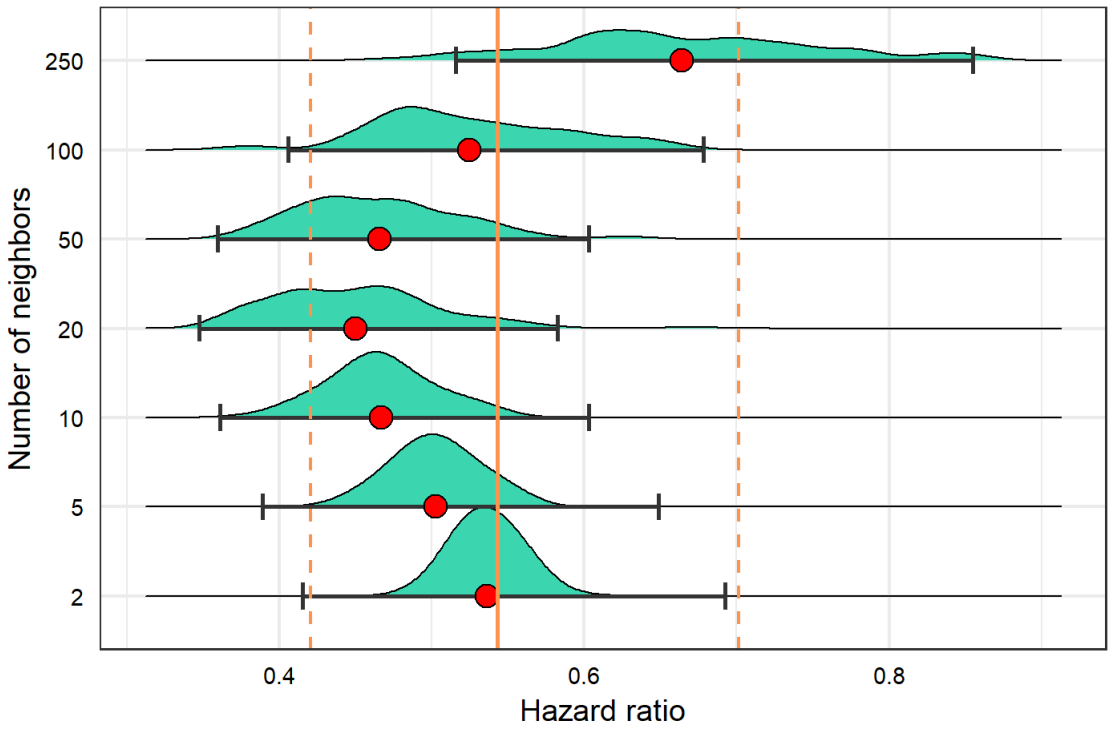
$k = 10$



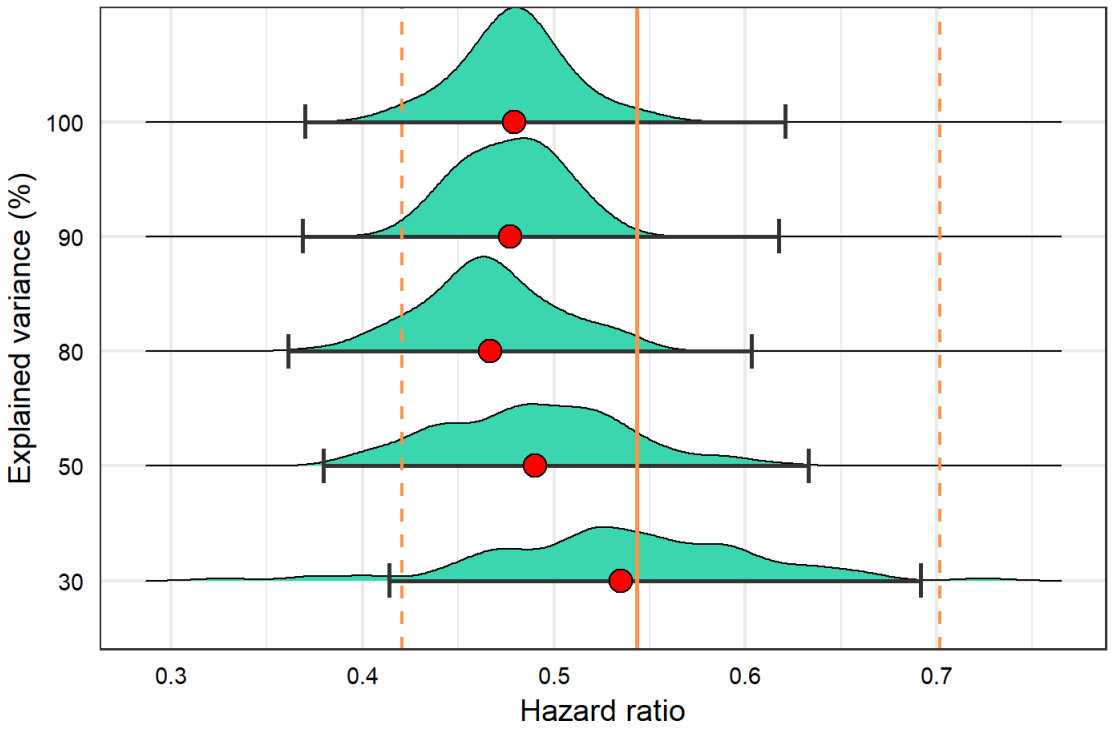
Avatars robustness – 100 synthetic datasets

- Arm 1 vs Placebo – Treatment effect on McDonald criterion (Cox)

$ncp = 8$
$\%var = 80$



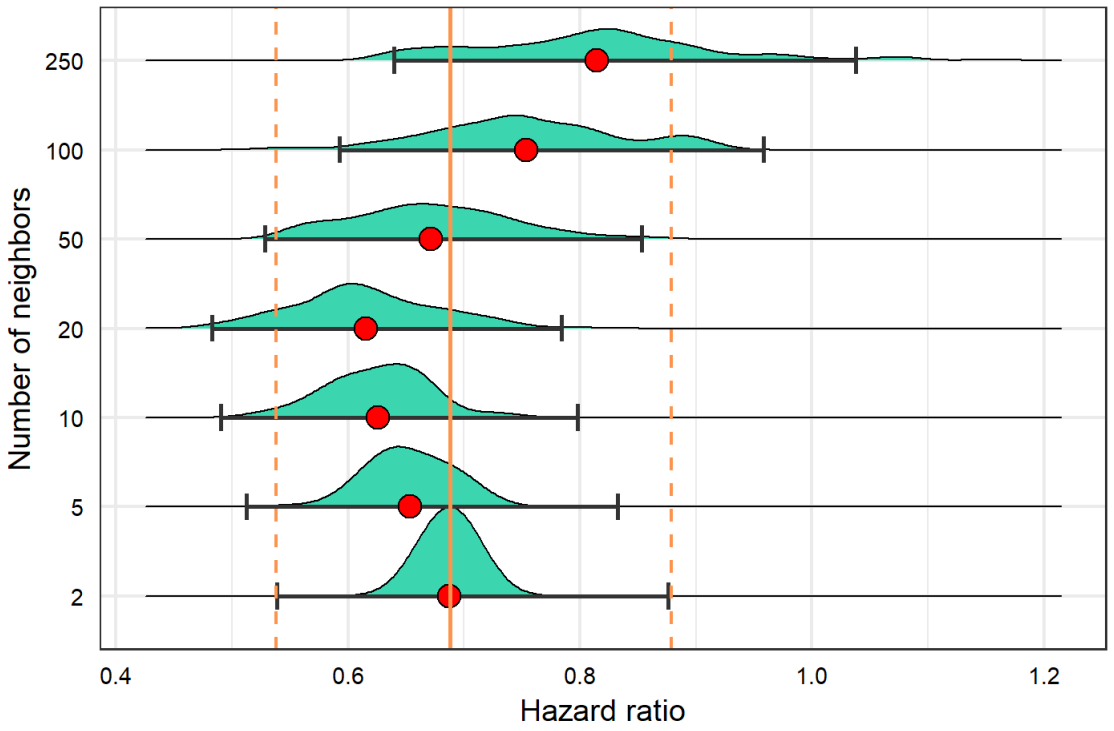
$k = 10$



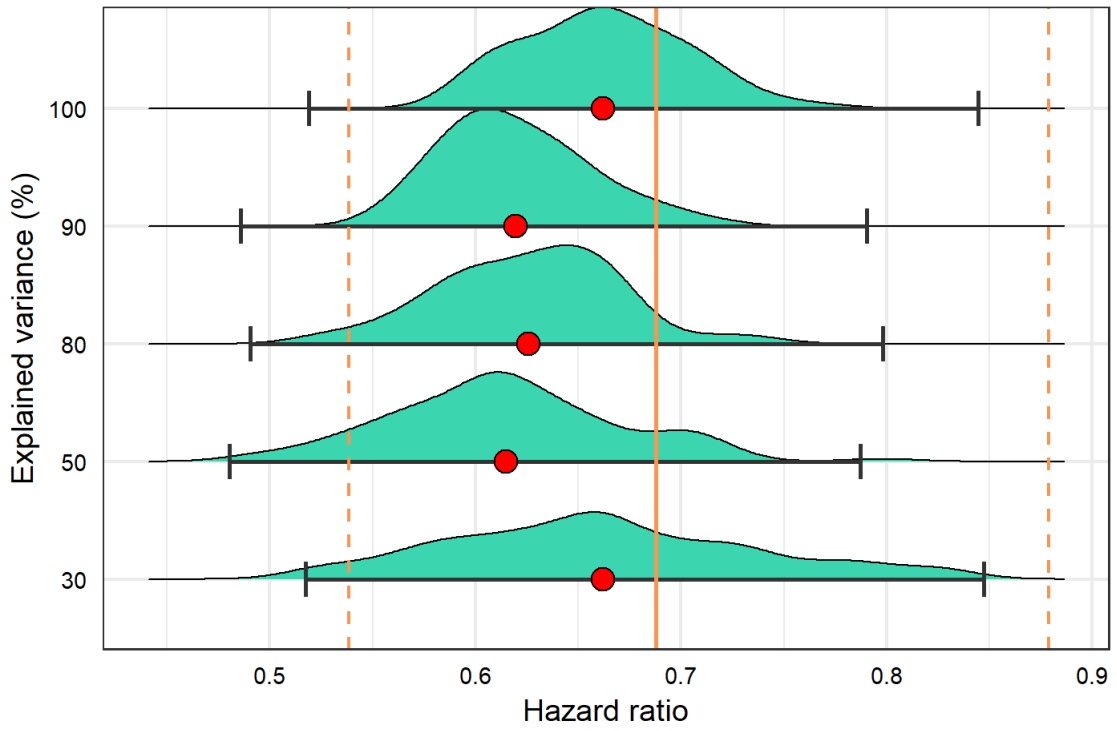
Avatars robustness – 100 synthetic datasets

- Arm 2 vs Placebo – Treatment effect on McDonald criterion (Cox)

$ncp = 8$
$\%var = 80$

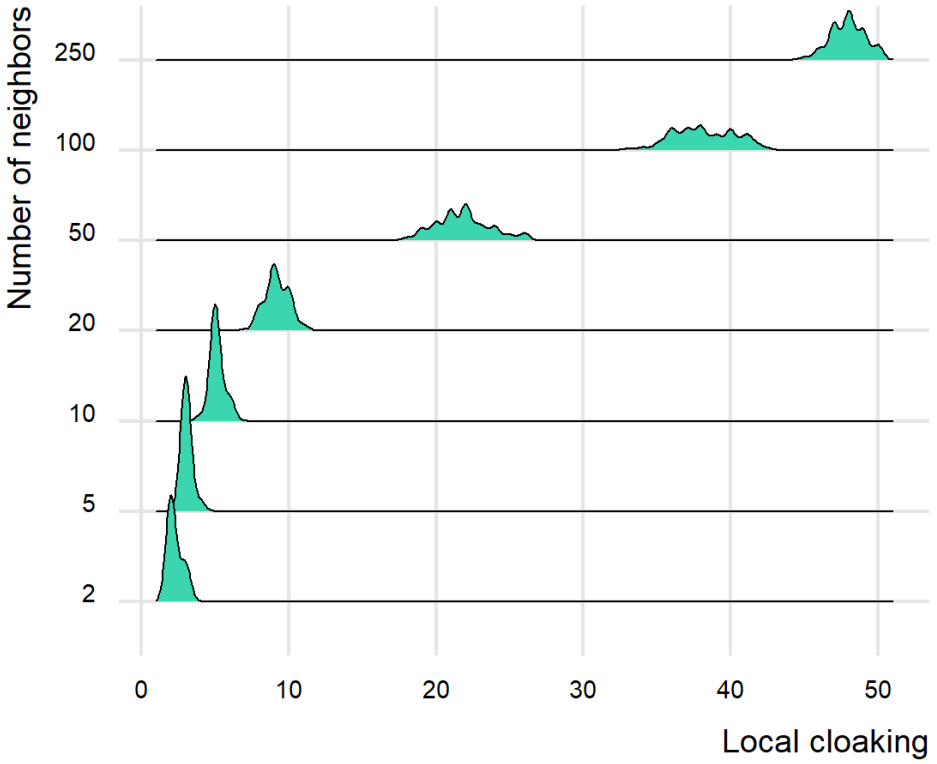


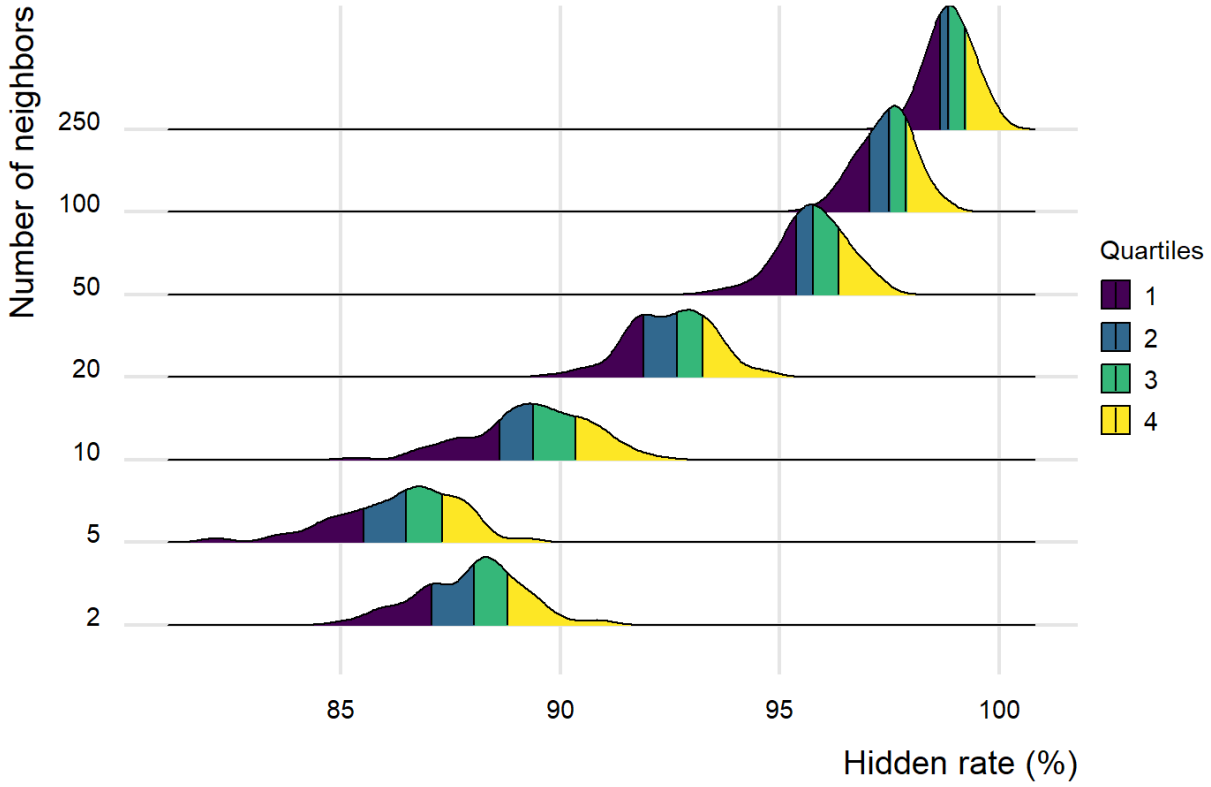
$k = 10$



Avatar method – Security metrics

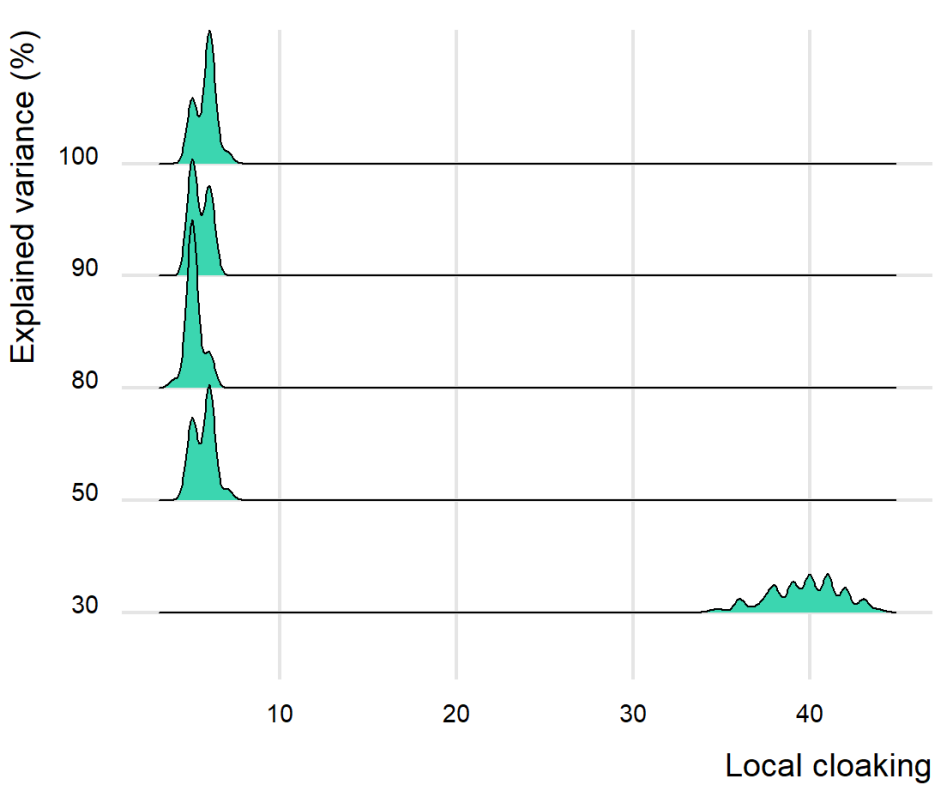
$ncp = 8$
$\%var = 80$

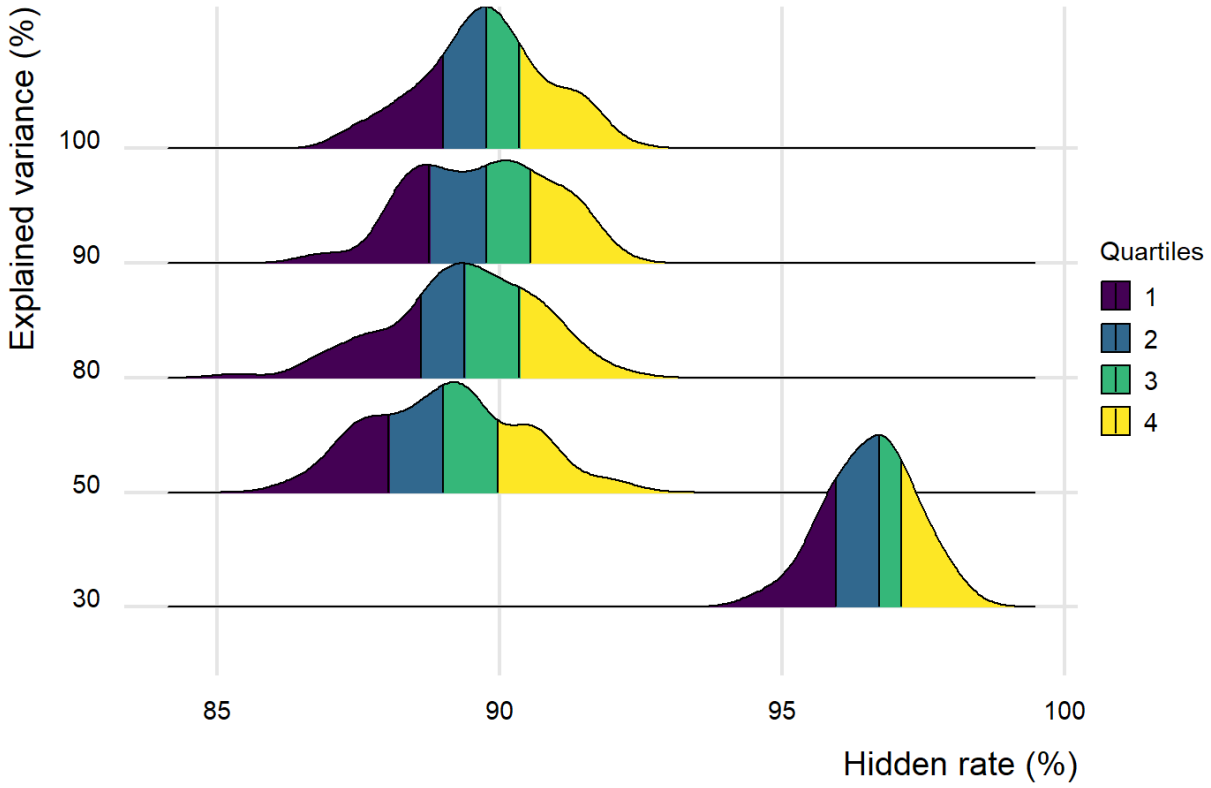




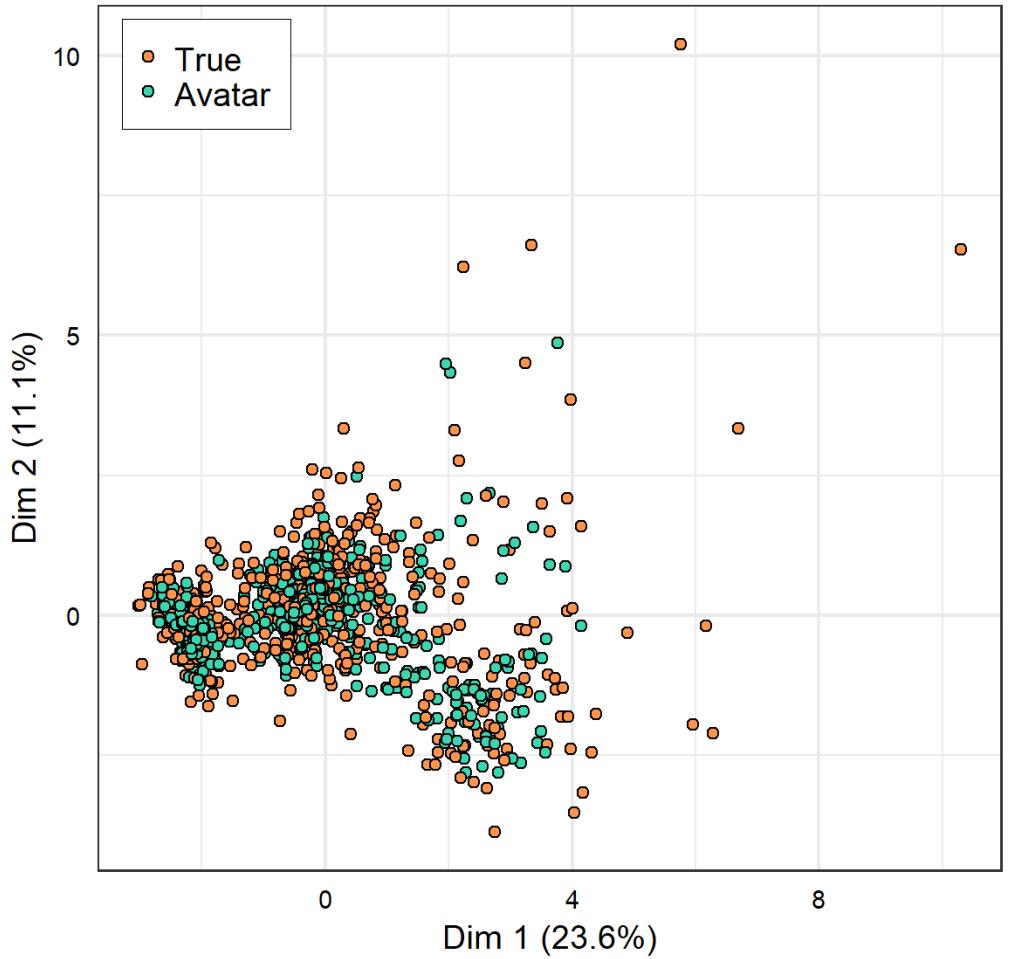
Avatar method – Security metrics

$k = 10$





PRIMUS selected avatar

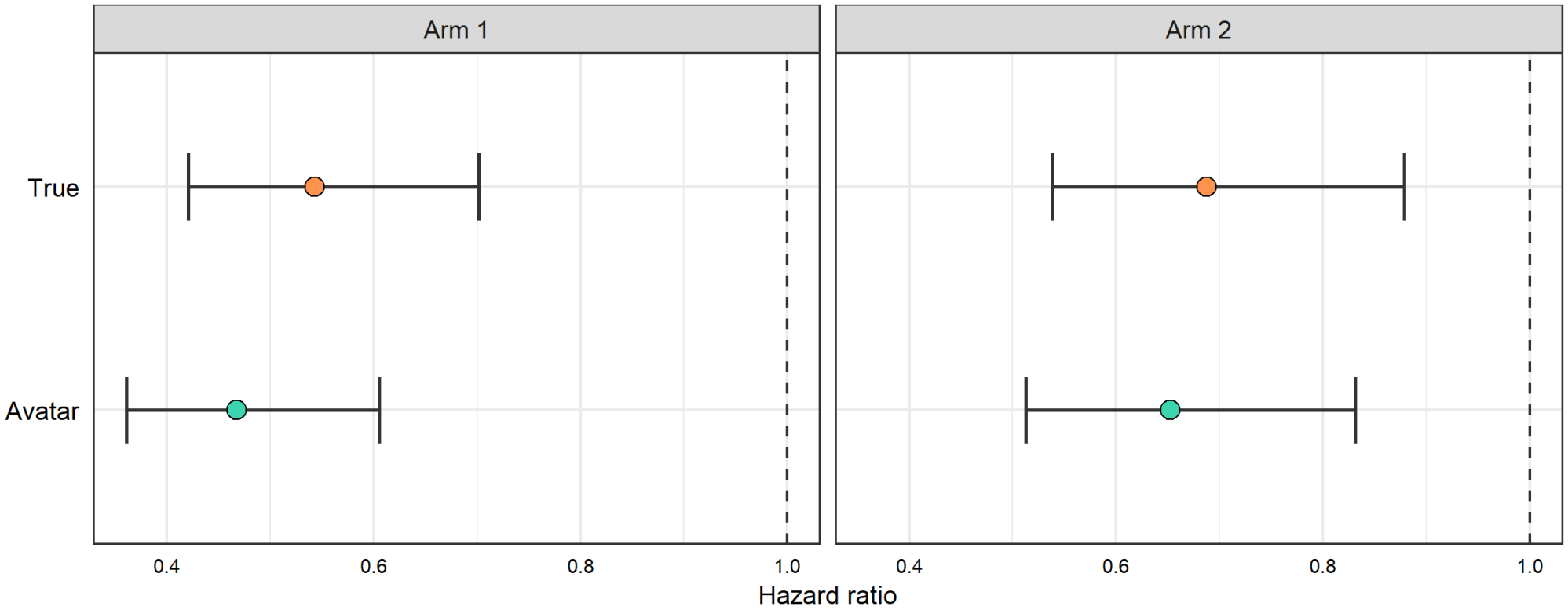


- $k = 10$
- $\%var = 80 \rightarrow ncp = 8$

- Hidden rate = 90%
- Local cloaking = 5

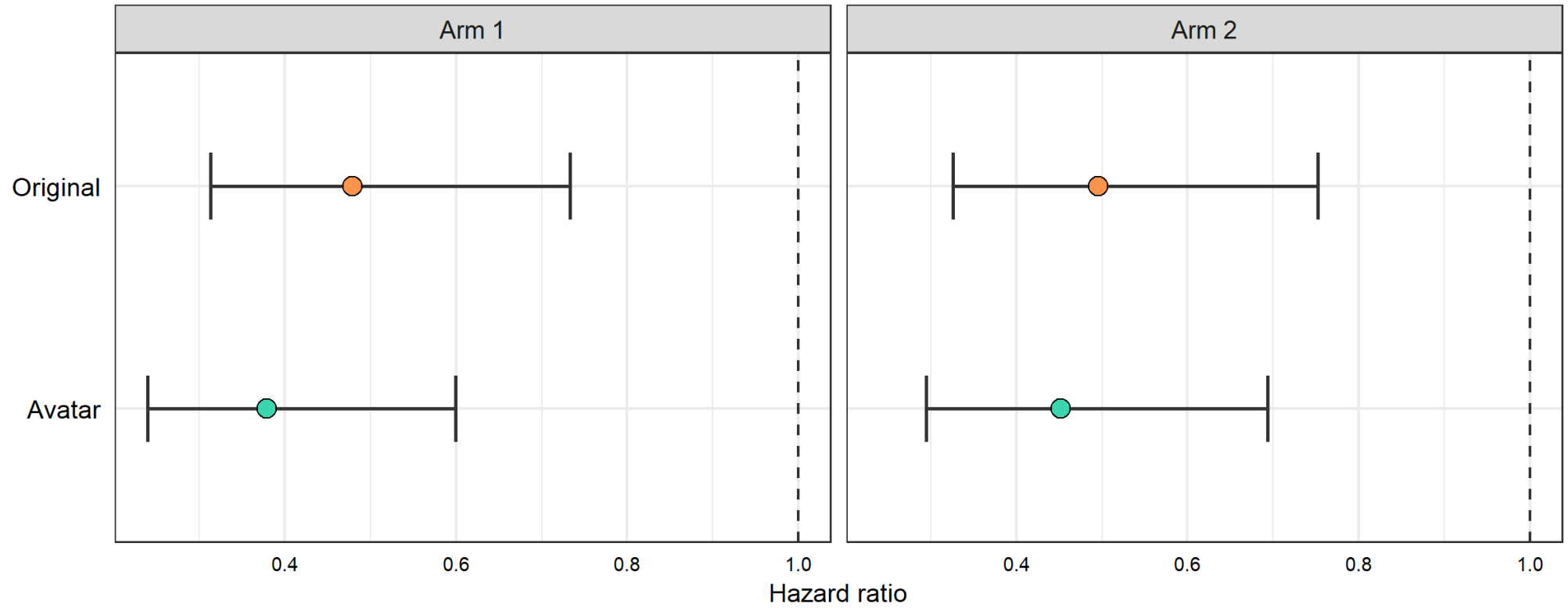
PRIMUS selected avatar

- Primary endpoint: McDonald criterion



PRIMUS selected avatar

- Secondary endpoint: clinically definite multiple sclerosis criterion



Conclusions & Perspectives

Conclusion

- No need to risk **re-identification** of patients
- **Avatar method** keeps global characteristics and protect patients

Conclusion

- No need to risk **re-identification** of patients
- **Avatar method** keeps global characteristics and protect patients

Data characteristics

Data security

Ongoing works & Perspectives

ICAN project – Intracranial aneurysm study

KTD-innov project – Kidney transplantation study

First year medical study grades database

Use of column weights parameter for multidimensional projection step

Acknowledgments

Team 3

Sophie Limou
Gilles Blancho
Pierre-Antoine Gourraud
Nicolas Vince
Fabienne Haspot
Sonia Bourguiba-Hachemi
Martin Morin
Venceslas Douillard
Léo Boussamet
Nayane Brito
Vincent Mauduit
Irène Charles
Axelle Durand
Rokhaya Ba
Sirine Sayadi
Chadia Ed Driouch
Stanislas Demuth
Lucas Béraud
Ketsia Mortant
Augustin Moreau
Julien Bouchet
Justine Baron
Hugo Barbot

Octopize

Olivier Breillacq
Mogan Guillaudeux
Julien Petot
Charles-Axel Dein
Tom Crasset
Gaël Russeil

Clinique des données

Matilde Karakachoff
Matthieu Wargny
Sandrine Coudol
Thomas Goronflot

CHU Rennes

Gilles Edan

Variables contribution