

Improving research transparency in Epidemiology via synthetic datasets when analysing sensitive or confidential data

Daniel Major-Smith

University of Bristol, UK

25th September 2024

Disclosure: No synthetic participants were harmed in the making of this presentation

WCE






WORLD CONGRESS OF EPIDEMIOLOGY 2024





METHOD ARTICLE

Releasing synthetic data from the Avon Longitudinal Study of Parents and Children (ALSPAC): Guidelines and applied examples [version 1; peer review: awaiting peer review]

Daniel Major-Smith ¹, Alex S. F. Kwong ^{1,2}, Nicholas J. Timpson ^{1,3},
Jon Heron ^{1,3}, Kate Northstone ¹

¹Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, England, BS8 2BN, UK

²Division of Psychiatry, The University of Edinburgh, Edinburgh, Scotland, EH10 5HF, UK

³MRC Integrative Epidemiology Unit, University of Bristol, Bristol, England, UK

Science is broken!

Science
Fictions

Stuart
Ritchie

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The Hardest Science

Everything is fucked: The syllabus

PUBLISHED ON *August 11, 2016*

PSY 607: Everything is Fucked

Prof. Sanjay Srivastava

Sanjay Srivastava



Exposing Fraud,
Bias, Negligence
and Hype in Science



<https://thehardestscience.com/2016/08/11/everything-is-fucked-the-syllabus/>

WCE

WORLD CONGRESS OF EPIDEMIOLOGY 2024



How to improve science?

- Need open data to see whether results are reproducible (Munafo et al., 2017; Smaldino et al., 2019)
 - Many papers don't openly share data
 - And even if data available, analyses sometimes not reproducible (Minocher et al., 2021)
 - Can help spot any errors in analyses
 - (Can also help readers test and understand new/advanced analysis methods)
- If data (and code, importantly!) are available, this increases trust in results and conclusions
 - Open data does not solve all problems, but is part of solution (including pre-registration/registered reports, better stats training, altering academic incentive structures, etc.)



Summary



Medical researchers do not commonly share their data or code. Although the number of researchers declaring that their data are publicly available is increasing, declared availability does not necessarily guarantee actual availability

Study design



Systematic review with meta-analysis of individual participant data

Data sources

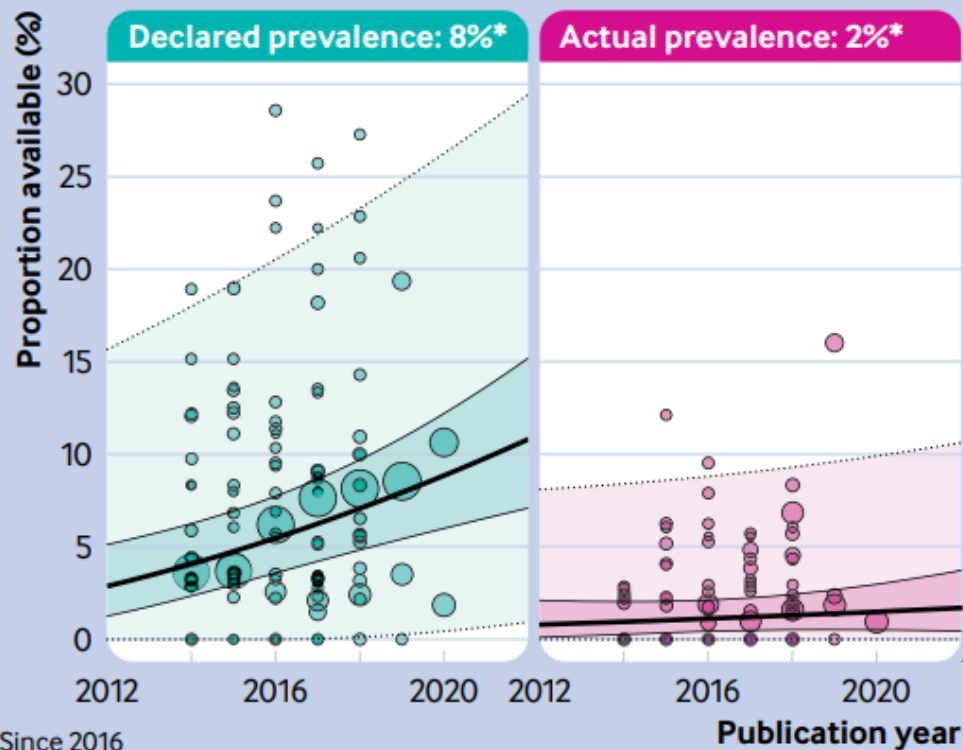


105 meta-research studies
2 121 580 medical publications



Risk of bias:
8% Low, 87% high, 6% unclear

Outcomes Declared availability v actual availability



Data sharing prevalence by publication year with fitted meta-regression lines

- 95% confidence interval
- 95% prediction interval
- Circles are scaled relative to the natural log of the sample size

Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data

Daniel G Hamilton,^{1,2} Kyungwan Hong,³ Hannah Fraser,¹ Anisa Rowhani-Farid,³ Fiona Fidler,^{1,4} Matthew J Page⁵

For public code sharing, both the prevalence of declared and actual availability were estimated to be <0.5% since 2016.



WCE

WORLD CONGRESS OF EPIDEMIOLOGY 2024



But sharing data is not always possible

- Potential concerns over:
 - 1) Data confidentiality/anonymity/sensitivity
 - 2) Ensuring that only legitimate researchers are able to access the resource
- Especially challenging for longitudinal population-based studies
 - E.g., ALSPAC has data on ~15,000 mothers, their partners and offspring, with over 100,000 variables in total
 - Hence why ALSPAC has policy of not allowing data to be released alongside published articles
- Policy is there for a reason, but is difficult to square with open science best practices of data sharing...

Synthetic data – A potential solution?

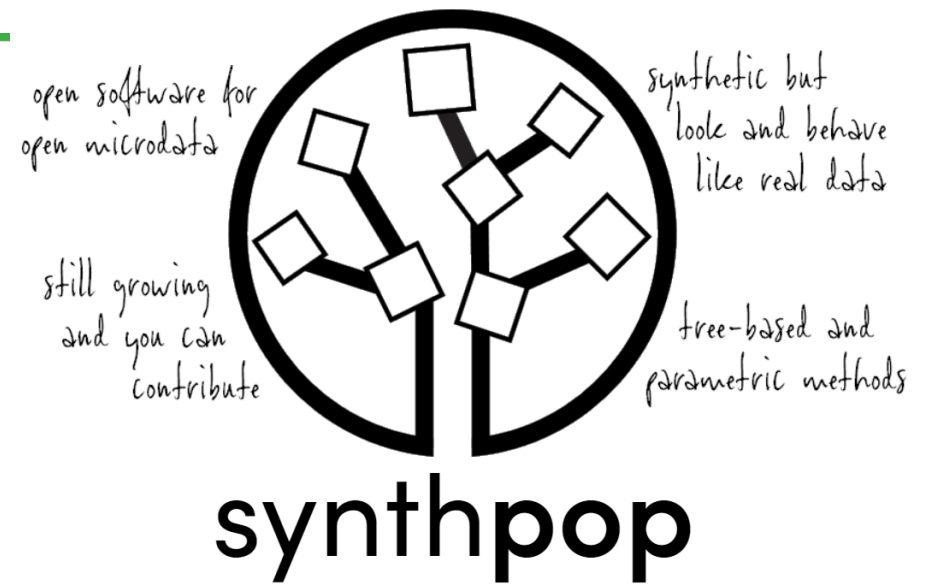
- Synthetic data are modelled on the original observed data
 - Maintains distributions of variables (means, variances, cell counts, etc.)
 - Maintains relationships between variables
 - But importantly observations are entirely simulated so do not correspond to real-life individuals
- Synthetic datasets keep the important features and structures of the observed data, while preserving anonymity
 - (Although correspondence between observed and synthetic data will not be perfect)

Synthetic data – A potential solution?

- Can release this synthetic data with the published paper so readers can:
 - Explore the raw (synthetic) data
 - Understand the analyses better
 - Reproduce analyses themselves
- While synthetic data will not be exactly the same as the observed data, it does add a further level of openness, accountability and transparency
 - Research is ‘quasi-reproducible’ (Shepherd et al., 2017)

'Synthpop' package

- R package 'synthpop' can be used to generate synthetic datasets (<https://www.synthpop.org.uk/>)
- Example based on open subset of ALSPAC data (<https://osf.io/8sgze>)
 - RQ: Are maternal postnatal depressive symptoms associated with offspring depression in adolescence?



'Synthpop' package

1. Synthesise data (default CART method)

```
dat_syn <- syn(dat, seed = 13327)
```

2. Remove uniquely-replicated individuals

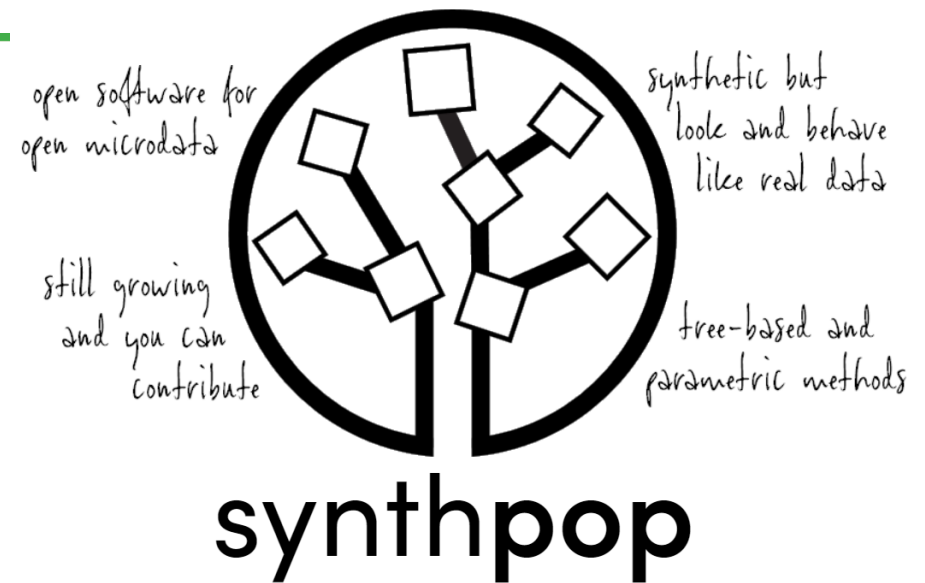
```
dat_syn <- sdc(dat_syn, dat,  
  rm.replicated.uniques = TRUE)
```

3. Compare variable distributions between observed and synthetic data

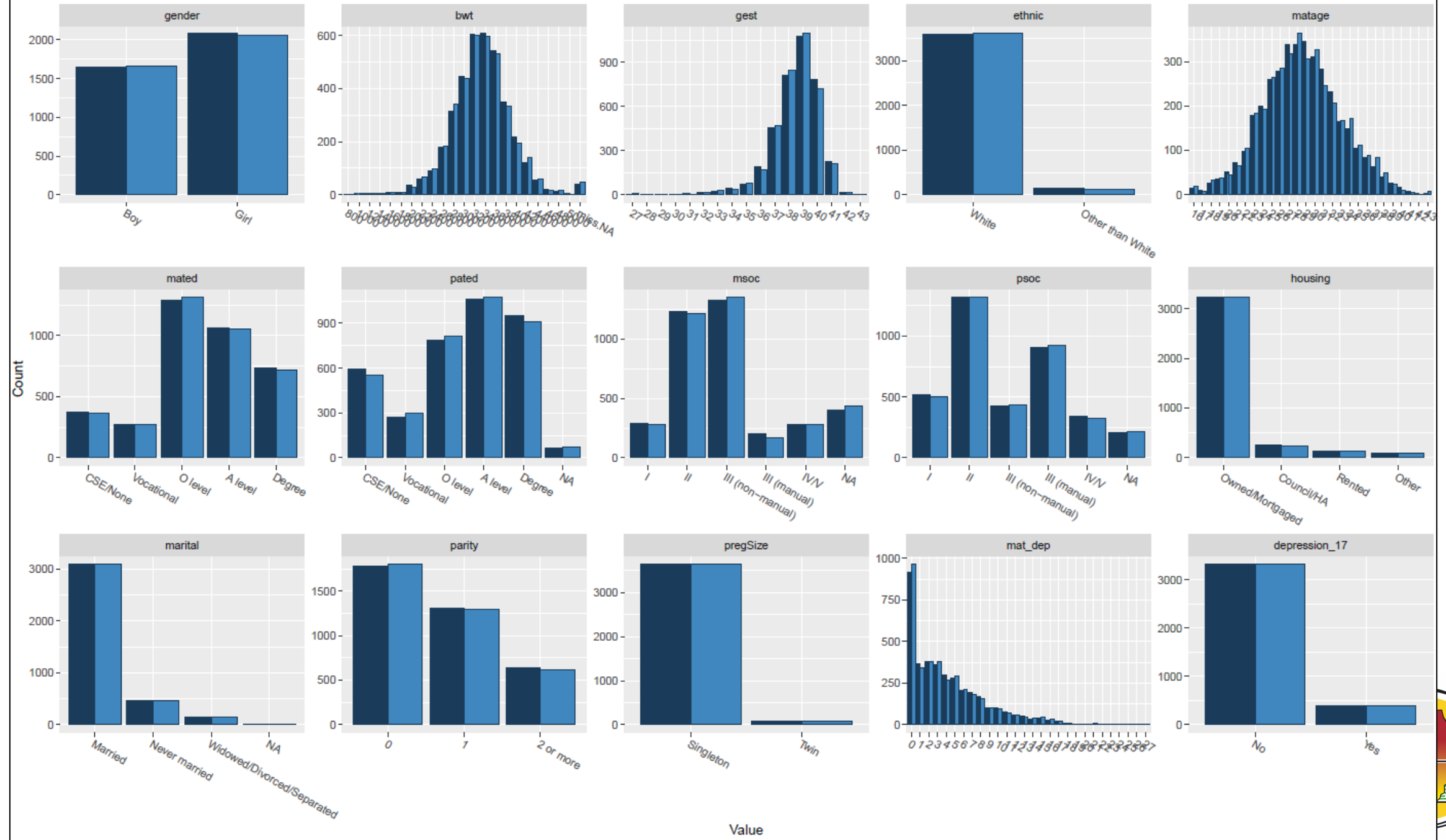
```
compare(dat_syn, dat, stat = "count")
```

4. Compare relationships between observed and synthetic data

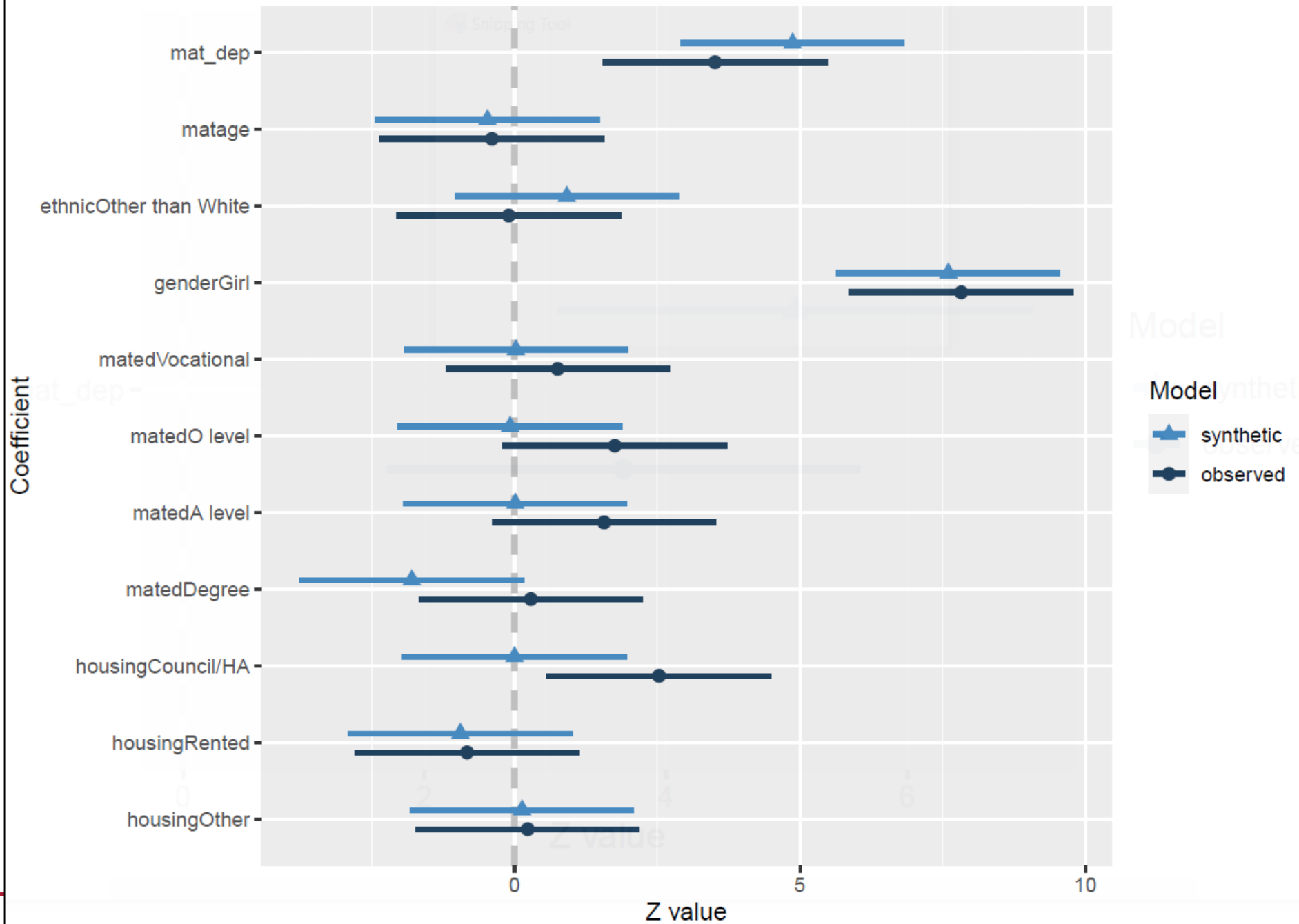
```
model_syn <- glm.synds(depression_17 ~ mat_dep + mat_age +  
  ethnic + gender + mated + housing,  
  family = "binomial", data = dat_syn)  
compare(model_syn, dat)
```



observed synthetic



Z values for fit to depression_17_17



Some caveats

- Synthetic datasets should *not* be used for research purposes
- Hopefully obvious, but release the original/observed data if possible
 - Synthetic data should be a last resort, not default!
- While non-disclosive, ask the original data owner about sharing synthetic data
 - E.g., co-developed guidelines for sharing synthetic ALSPAC data
- ‘Synthpop’ designed for datasets with independent observations
 - May not be suitable (or work as well) for more complex datasets (e.g., hierarchical/multi-level data, social networks)



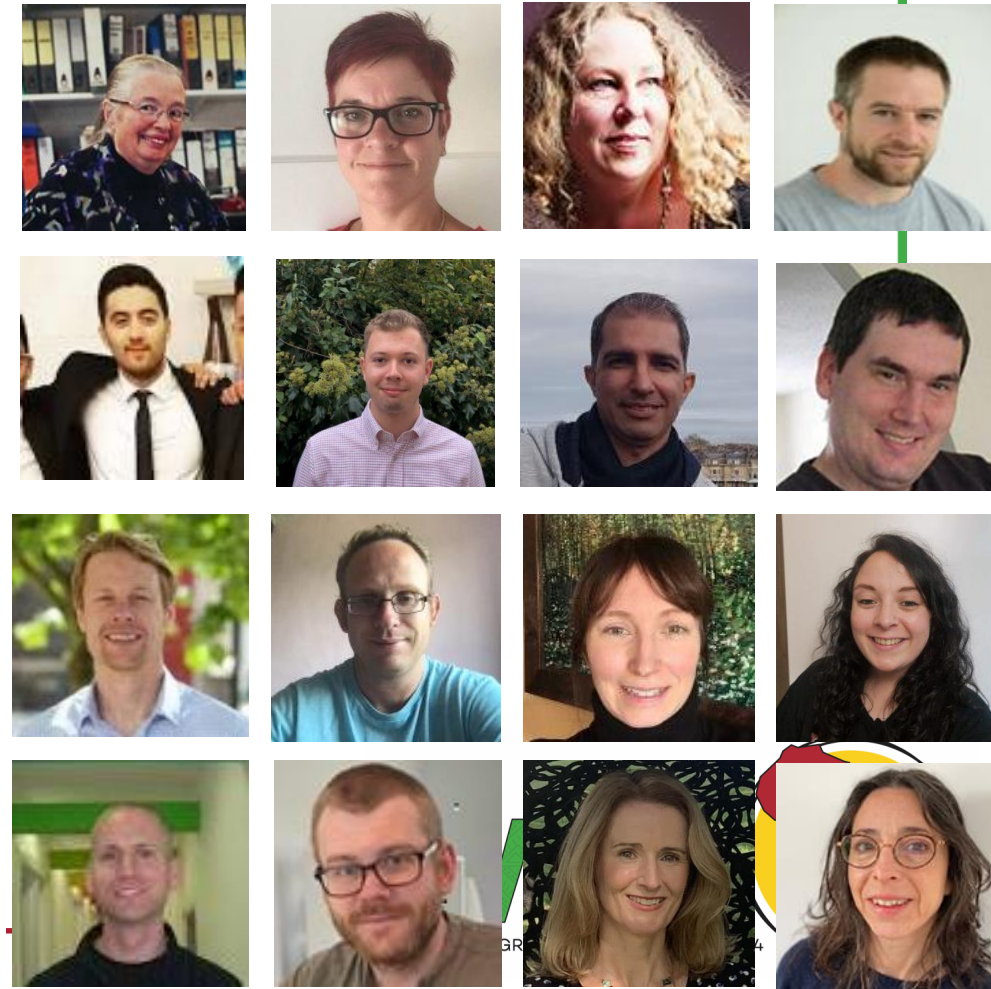
Take-home message

- In summary: Share your data! (but if you can't, share synthetic data)

- Thanks:
 - All the wonderful co-authors and collaborators on the Beliefs, Behaviours and Health Research Group
 - John Templeton Foundation for funding
 - ALSPAC for having such amazing participants & data!

- More information:
 - Paper in *Wellcome Open Research* (<https://wellcomeopenresearch.org/articles/9-57/v1>)
 - GitHub repo for code and synthetic data (<https://github.com/djsmith-90/synthetic-data>)
 - (Plus other repos with synthetic ALSPAC data)
 - Openly-available subset of ALSPAC data (<https://osf.io/8sgze>)

- Contact:
 - Email: dan.smith@bristol.ac.uk
 - Bluesky: [@djsmith90.bsky.social](https://bsky.app/profile/@djsmith90.bsky.social)



References/Useful links

- Major-Smith, D., Kwong, A. S., Timpson, N. J., Heron, J., & Northstone, K. (2024). Releasing synthetic data from the Avon Longitudinal Study of Parents and Children (ALSPAC): Guidelines and applied examples. Wellcome Open Research, 9, 57. (<https://wellcomeopenresearch.org/articles/9-57>)
- Nowok, B., Raab, G.M. & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. J. Stat. Softw., 74. (<https://www.jstatsoft.org/article/view/v074i11/0> – ‘Synthpop’ website also very useful: <https://www.synthpop.org.uk/>)
- Raab, G.M., Nowok, B. & Dibben, C. (2017). Guidelines for Producing Useful Synthetic Data. arXiv Prepr. (<https://arxiv.org/abs/1712.04078>)
- Raghunathan, T.E. (2021). Synthetic data. Annu. Rev. Stat, 8, 129–140. (<https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-040720-031848>)
- Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie Du Sert, N., et al. (2017). A manifesto for reproducible science. Nat. Hum. Behav., 1, 1–9. (<https://www.nature.com/articles/s41562-016-0021>)
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. Royal Society open science, 3(9), 160384. (<https://royalsocietypublishing.org/doi/full/10.1098/rsos.160384>)
- Minocher, R., Atmaca, S., Bavero, C., McElreath, R., & Beheim, B. (2021). Estimating the reproducibility of social learning research published between 1955 and 2018. Royal Society Open Science, 8(9), 210450. (<https://royalsocietypublishing.org/doi/full/10.1098/rsos.210450>)
- Hamilton, D. G., Hong, K., Fraser, H., Rowhani-Farid, A., Fidler, F., & Page, M. J. (2023). Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. bmj, 382. (<https://doi.org/10.1136/bmj-2023-075767>)
- Shepherd, B. E., Blevins Peratikos, M., Rebeiro, P. F., Duda, S. N., & McGowan, C. C. (2017). A pragmatic approach for reproducible research with sensitive data. American Journal of Epidemiology, 186(4), 387-392. (<https://doi.org/10.1093/aje/kwx066>)
- Mathur, M. B., & Fox, M. P. (2023). Toward open and reproducible epidemiology. American Journal of Epidemiology, 192(4), 658-664. (<https://doi.org/10.1093/aje/kwad007>)

