University of BRISTOL

ALSPAC
Bristol Medical School
Oakfield House, Oakfield Grove
Bristol, BS8 2BN

www.alspac.bristol.ac.uk

CHECKLIST FOR GENERATING SYNTHETIC DATASETS FROM THE ALSPAC RESOURCE USING ‘synthpop’: v1. 28th November 2023

Any projects using ALSPAC data that generate a synthetic dataset must complete this checklist before the synthetic dataset is made openly available. **We expect to process all requests to deposit synthetic datasets within two weeks of receipt of a completed checklist.** As with any papers written that use ALSPAC data, we will provide advice and feedback to authors where we feel this may be helpful. Please send the completed checklist, dataset and associated scripts and publication to alspac-exec@bris.ac.uk *prior* to making your synthetic dataset available and allow sufficient time for processing. You may wish to submit this form at the same time as your publication checklist. Our review will not necessarily check all the points that you have ticked as this is not practical, however, by ticking the boxes below, you take full responsibility for ensuring these are correct.

Name of corresponding author:

ALSPAC Data Buddy (if applicable):

Title of paper for which the synthetic data was generated:

Proposal/B number: **Paper/C number (if applicable):**

	Yes	No
1. I have reduced the number of variables and observations in the dataset I have synthesised to only those required to replicate the results in the paper (e.g., if the original dataset contains 15,645 observations and 20 variables, while the final analyses contain 4,000 observations and only use 10 variables, all additional observations and variables have been removed prior to synthesis).		
2. I confirm that the synthetic dataset I have generated contains less than 50 variables [if you need to synthesise more than 50 variables, please talk to ALSPAC and provide justification first].		
3. I can confirm that no individuals are uniquely identified in both the observed and synthetic datasets.		
4. I have undertaken a manual check to ensure there are no unique replicates in either the observed or synthetic datasets.		
5. I confirm that the distributions of all synthesised variables in the synthetic dataset are similar to those in the observed data.		
6. I have checked the relationships between synthesised key variables are comparable to those in the observed data (e.g., via univariable and multivariable regressions).		
7. I have included a variable at the beginning of the synthetic dataset named ‘FALSE_DATA’, with values of ‘FALSE_DATA’ for all observations, to ensure it is clear that the dataset contains synthetic data, rather than real observations (see Nowok *et al.* 2017).		

8. I have included a disclaimer alongside the synthetic data (and in the paper to be published if applicable), making it clear to users that the data are synthetic and should not be used for any subsequent research or publications [see footnote 1].		
9. I agree to provide details of any downloads/requests to use the synthetic data if ALSPAC request such information (where it is possible).		
10. I will publish the script that developed the synthetic dataset to sit alongside the dataset (including code for all variable name changes and variable recodes and derivations from the original ALSPAC data, to facilitate reproducibility).		
11. I will provide the DOI or a suitable weblink as to where the synthetic data are stored (e.g., GitHub (https://github.com/), Dryad (https://datadryad.org/stash), or Open Science Framework (https://osf.io/). Please provide here if you have it prepared but not yet made publicly available:		

Signature: **Date:**

FOOTNOTES

1. Disclaimer

We recommend the following disclaimer:

"*These are synthesised ALSPAC datasets, and are* not *suitable for research purposes. The relations between variables are unlikely to be maintained perfectly, so there is the risk when using these synthesised datasets that results may differ from the true data. Only the actual, observed, ALSPAC data should be used for formal research and analyses reported in published work.*".

References

Nowok, B., Raab, G.M. & Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Stat. J. IAOS*, 33, 785–796.