Longitudinal and Life Course Studies • vol 13 • no 2 • 335–341
© Authors 2022 • Paper received October 2021/Published online Feburary 2022
Online ISSN 1757-9597 • https://doi.org/10.1332/175795921X16428748347208

CORRESPONDENCE

Letter to the editor: Don't forget survey data: 'healthy cohorts' are 'real-world' relevant if missing data are handled appropriately

Richard J. Silverwood, r.silverwood@ucl.ac.uk
Alissa Goodman, alissa.goodman@ucl.ac.uk
George B. Ploubidis, g.ploubidis@ucl.ac.uk
University College London, UK

To cite this article: Silverwood, R., Goodman, A. and Ploubidis, G. (2022) Letter to the editor: Don't forget survey data: 'healthy cohorts' are 'real-world' relevant if missing data are handled appropriately, *Longitudinal and Life Course Studies*, vol 13, no 2, 335–341, DOI: 10.1332/175795921X16428748347208

Dear Professor Joshi,

We write to you regarding the published article 'Are "healthy cohorts" real-world relevant? Comparing the National Child Development Study (NCDS) with the ONS Longitudinal Study (LS)' by Archer et al (2020). The authors report that NCDS is unrepresentative of age-matched LS respondents, but that despite differences in sample characteristics, longitudinal associations were similar in the NCDS and LS samples. They attribute the discrepancy between NCDS and LS to a 'healthy cohort' effect and propose that creating non-response weights from administrative data should be used. While we agree with Archer et al that administrative data have the potential to inform missing data analyses in longitudinal surveys, the authors omit to mention that even without administrative data there are already methods available to researchers to restore sample representativeness using survey information alone that have been shown to be highly effective.

To demonstrate the effectiveness of using survey information – without augmentation by administrative data – in restoring sample representativeness in NCDS with respect to the LS, we present Table 1 from their manuscript, with additional columns from our own analyses. We accounted for non-response at age 46 and 55 with multiple imputation (MI), using chained equations (Azur et al, 2011; White et al, 2011; Harel et al, 2018) to generate 50 imputed datasets.[1] The imputation phase included 'auxiliary variables' (Carpenter and Kenward 2012) from earlier sweeps of

Table 1: Sample characteristics (prevalence and 95% confidence interval unless otherwise stated)

	NCDS 2004 (age 46) n = 8,689[a]			ONS LS 2001 (age 45)		NCDS 2013 (age 55) n = 8,107[a]			ONS LS 2011 (age 55)	
	Archer et al Table 1	Our calculations		Archer et al Table 1 n = 7,157[c]	Archer et al Table S4 n = 6,393[d]	Archer et al Table 1	Our calculations		Archer et al Table 1 n = 7,052[c]	Archer et al Table S4 n = 6,170[d]
		Observed data	MI[b]				Observed data	MI[b]		
Long-term limiting illness										
Yes				14.9	15.0	19.7	19.7 (18.8, 20.6)	22.6 (21.5, 23.6)	22.8	22.5
No				85.1	85.0	80.3	80.3 (79.4, 81.2)	77.4 (76.4, 78.5)	77.2	77.5
Missing (n)				141	99	115	115	0	155	127
Sex										
Male	48.7	48.8 (47.7, 49.8)	51.1 (50.2, 51.9)	49.4	49.9	48.5	48.5 (47.4, 49.6)	50.7 (49.9, 51.6)	49.3	49.9
Female	51.3	51.2 (50.2, 52.3)	48.9 (48.1, 49.8)	50.6	50.1	51.5	51.5 (50.4, 52.6)	49.3 (48.4, 50.1)	50.7	50.1
Missing (n)	0	0	0	0	0	0	0	0	0	0
Ethnicity										
White	98.0	98.1 (97.8, 98.3)	96.8 (96.5, 97.1)	90.3	96.9	97.9	97.9 (97.6, 98.2)	96.8 (96.4, 97.1)	88.3	95.8
Non-white	2.0	1.9 (1.7, 2.2)	3.2 (2.9, 3.5)	9.7	3.1	2.1	2.1 (1.8, 2.4)	3.2 (2.9, 3.6)	11.7	4.2
Missing (n)	0	10	0	113	113	0	7	0	116	95
Region										
South	47.9	47.9 (46.9, 49.0)	46.3 (45.3, 47.2)	49.4	47.2	46.0	48.1 (47.0, 49.2)	45.8 (44.9, 46.8)	50.1	47.4
North	46.1	46.1 (45.1, 47.2)	47.0 (46.0, 47.9)	45.3	47.0	48.1	46.0 (44.9, 47.1)	47.8 (46.8, 48.8)	44.6	46.7
Wales	6.0	6.0 (5.5, 6.5)	6.8 (6.2, 7.3)	5.3	5.8	6.0	6.0 (5.5, 6.5)	6.4 (5.9, 6.9)	5.3	5.9
Missing (n)	3	3	0	2	2	0	0	0	0	0
Employment status										
Full-time	69.0	69.0 (68.0, 70.0)	66.3 (65.3, 67.3)	61.1	62.4	61.2	61.2 (60.2, 62.3)	57.8 (56.5, 59.0)	55.2	56.0
Part-time	18.4	18.3 (17.5, 19.2)	17.0 (16.2, 17.8)	17.7	18.0	20.2	20.2 (19.3, 21.1)	18.8 (17.9, 19.7)	19.0	19.5

(*Continued*)

Table 1: (*Continued*)

	NCDS 2004 (age 46) n = 8,689[a]			ONS LS 2001 (age 45)		NCDS 2013 (age 55) n = 8,107[a]			ONS LS 2011 (age 55)	
	Archer et al Table 1	Our calculations		Archer et al Table 1 n = 7,157[c]	Archer et al Table S4 n = 6,393[d]	Archer et al Table 1	Our calculations		Archer et al Table 1 n = 7,052[c]	Archer et al Table S4 n = 6,170[d]
		Observed data	MI[b]				Observed data	MI[b]		
Unemployed	1.7	1.6 (1.4, 1.9)	2.4 (2.0, 2.8)	3.2	3.1	2.9	2.9 (2.5, 3.2)	4.0 (3.5, 4.6)	4.3	4.3
Long-term sick/disabled	4.0	4.0 (3.6, 4.4)	6.1 (5.5, 6.6)	6.3	6.4	5.2	5.2 (4.7, 5.7)	7.6 (6.9, 8.3)	9.2	9.1
Looking after home/family	5.4	5.4 (4.9, 5.9)	6.0 (5.5, 6.5)	7.3	6.2	6.2	6.2 (5.7, 6.8)	7.2 (6.5, 7.8)	5.1	4.5
Other[e]	1.7	1.7 (1.4, 2.0)	2.1 (1.7, 2.5)	4.4	4.0	4.3	4.3 (3.9, 4.8)	4.6 (4.1, 5.2)	7.1	6.6
Missing (n)	0	0	0	3	2	120	120	0	151	126
Social class NS-SEC										
Professional/ higher management	41.9	41.9 (40.8, 42.9)	[g]	33.9	34.8	35.7	35.7 (34.7, 36.8)	[g]	29.2	30.0
Intermediate	19.8	19.8 (19.0, 20.7)	[g]	18.4	18.6	23.1	23.1 (22.2, 24.0)	[g]	20.1	20.7
Routine and manual	25.7	25.7 (24.7, 26.6)	[g]	28.0	28.3	20.9	20.9 (20.0, 21.8)	[g]	25.1	25.0
Other[f]	12.7	12.7 (12.0, 13.4)	[g]	19.7	18.3	20.3	20.3 (19.4, 21.2)	[g]	25.6	24.3
Missing (n)	27	27	[g]	3	2	120	187	[g]	0	0
Marital status										
Married	71.1	71.1 (70.1, 72.0)	67.3 (66.3, 68.3)	68.7	67.8	71.5	71.5 (70.5, 72.5)	65.6 (64.4, 66.8)	70.0	69.3
Divorced/ separated/ widowed	17.5	17.5 (16.7, 18.3)	19.8 (18.9, 20.7)	18.7	19.1	18.6	18.6 (17.8, 19.5)	21.9 (20.8, 22.9)	19.4	19.5
Single	11.5	11.5 (10.8, 12.2)	12.9 (12.2, 13.7)	12.7	13.1	9.9	9.9 (9.2, 10.5)	12.6 (11.8, 13.3)	10.6	11.2
Missing (n)	18	18	0	17	14	5	5	0	55	43

(*Continued*)

Table 1: (*Continued*)

	NCDS 2004 (age 46) n = 8,689[a]			ONS LS 2001 (age 45)		NCDS 2013 (age 55) n = 8,107[a]			ONS LS 2011 (age 55)	
	Archer et al Table 1	Our calculations		Archer et al Table 1 n = 7,157[c]	Archer et al Table S4 n = 6,393[d]	Archer et al Table 1	Our calculations		Archer et al Table 1 n = 7,052[c]	Archer et al Table S4 n = 6,170[d]
		Observed data	MI[b]				Observed data	MI[b]		
Living arrangements										
No partner				22.9	22.9	21.0	21.0 (20.1, 21.9)	25.8 (24.7, 27.0)	26.9	26.2
Spouse				68.1	67.7	69.1	69.1 (68.1, 70.1)	62.6 (61.4, 63.9)	64.9	65.1
Co-habiting				9.0	9.4	10.0	10.0 (9.3, 10.6)	11.5 (10.8, 12.3)	8.2	8.7
Missing (n)				47	41	0	0	0	49	36
Housing tenure										
Own – outright	14.3	14.3 (13.6, 15.1)	14.1 (13.4, 14.9)	16.2	15.7				34.0	35.3
Own – mortgage	71.5	71.5 (70.5, 72.4)	66.8 (65.8, 67.8)	63.5	65.1				43.4	43.9
Rent/other	14.2	14.2 (13.5, 15.0)	19.1 (18.2, 19.9)	20.3	19.2				22.6	20.8
Missing (n)	39	39	0	193	140				101	76

Notes:

[a] NCDS sample restricted to those resident in England and Wales.

[b] Multiple imputation. Imputation model includes analysis variables (with the exception of social class NS-SEC), predictors of non-response at sweep 7/9 and selected variables predictive of analysis variables.

[c] Including all LS respondents.

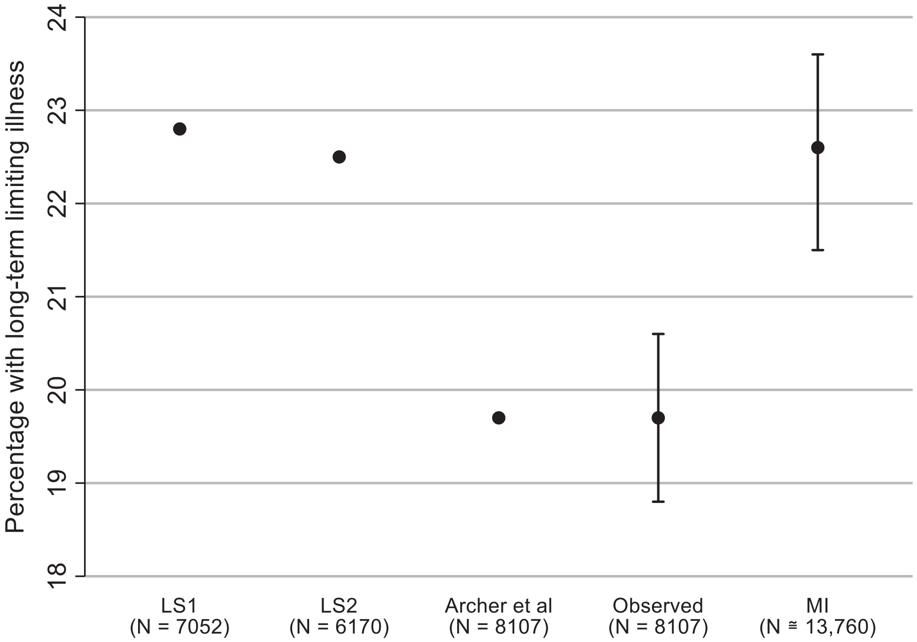
[d] Excluding LS respondents who arrived in the UK after age 16.

[e] Full-time education, government training scheme, retired, temporarily sick or disabled.

[f] Never worked, long-term unemployed, not working, unclassifiable.

[g] Social class NS-SEC not included in MI analysis due to collinearity with employment status.

Figure 1: Estimated prevalence of long-term limiting illness



Notes:

LS1: Estimate from ONS LS data including all LS respondents (from Archer et al Table 1).

LS2: Estimate from ONS LS data excluding LS respondents who arrived in the UK after age 16 (from Archer et al Table S4).

Archer et al: Estimate using observed NCDS Sweep 9 data (from Archer et al Table 1).

Observed: Estimate using observed NCDS Sweep 9 data (our own calculation).

MI: Estimate using multiple imputation (our own calculation).

NCDS that were associated with non-response at ages 46 and 55 and the outcome of interest (long-term limiting illness for example), as well as variables that are known to be associated only with the outcome of interest.[2]

In Table 1 we see that after accounting for loss to follow up with MI that includes auxiliary information from the NCDS survey itself, most estimates from NCDS are closer to those from LS, and do not show the discrepancy highlighted in the comparisons made by Archer et al. Results for the estimated prevalence of long-term limiting illness are shown in Figure 1. Taking into consideration that there are likely to be other potential sources of variation between NCDS and LS that were not accounted for by Archer et al that mean that we would not expect there to be a perfect match (age and calendar period effects, missing data handling in LS, minor differences in the way some questions were asked, and potential mode effects), our results suggest that using the methods described, NCDS sample representativeness with respect to LS was quite effectively restored.

These corrections do not constitute a formal test for missing data generating mechanisms, and there could be other variables in NCDS where we wouldn't be able to replicate the known population distribution with these methods. However, in our published work (Mostafa et al, 2021), we show that we are also able to replicate the known population distribution of educational attainment and marital status at age 50 based on external benchmarks (using the ONS Annual Population and Labour Force Surveys), as well as using internal benchmarks, by replicating the original distribution of paternal social class observed at the birth survey, and the distribution of cognitive ability at age 7.

While we have no doubt that the addition of information from population administrative data, in creation of weights, or by using these in multiple imputation or full information maximum likelihood could enhance these methods yet further, the extent of their benefits remains an open empirical question, and is likely to be modest relative to the survey data corrections described earlier. Our work in progress funded by the Economic and Social Research Council and Administrative Data Research UK (grant number ES/V006037/1) is augmenting these corrections using additional population administrative data, from hospital and educational records, and will be published in due course.

By making no attempt in their analyses to use survey responses to correct for missing data due to non-response/loss to follow up, Archer's et al findings are open to a clear misinterpretation by readers that there is nothing to be done to restore representativeness in NCDS and/or other longitudinal surveys, if administrative data are not used. This is far from the truth. Using appropriate methods, estimates from NCDS are indeed 'real-world' relevant and can be used for policy inference. Further guidance on how users can adopt these methods for missing data handling in NCDS in their own analyses is available in the NCDS Missing Data User Guide, and we also offer a programme of regular user training.[3]

Notes

[1] In this approach we view missing data analysis as an attempt to restore sample representativeness with respect to a well-defined target population. The target population of NCDS, and any other longitudinal survey, is dynamic, as changes occur for example due to mortality. Considering that the NCDS mortality rate is representative of the population (Mostafa et al, 2021), the target population in each sweep of NCDS needs to be adjusted accordingly to reflect these changes. In this instance the target population for our analyses are those born in Britain in 1958, alive at the time of data collection and still residing in Britain.

Missing values of the analysis variables were imputed using MI, with the exception of two variables: sex and ethnicity. We know sex (for all cohort members) and ethnicity (for virtually all cohort members) from previous sweeps. We therefore (singly) imputed these variables with their known values. We acknowledge that self-reported sex and ethnicity may vary over time within individuals, whereas this approach treats them as being fixed, but we would suggest that in 'real-world' analyses most analysts would be willing to make this assumption in order to handle missing data. After imputing these variables with their known values, sex is complete but ethnicity still has some missing values, which were handled using MI.

[2] Analyses of age 46 outcomes included 23 predictors of non-response at age 46 (as identified in Mostafa et al, 2021) and 11 variables considered predictive of underlying missing values: region at ages 0, 23 and 42, marital status at ages 23, 33 and 42, housing tenure at ages 23, 33 and 43, and employment status at ages 33 and 42. Analyses of age 55 outcomes included 30 predictors of non-response at age 55 (as identified in Mostafa et al, 2021) and 12 variables considered predictive of underlying missing values: region at ages 0, and 23, long-term limiting illness at ages 33 and 42, employment status at ages

33 and 50, marital status at ages 33, 42, 46 and 50, and living arrangements at ages 46 and 50.

[3] https://cls.ucl.ac.uk/data-access-training/handling-missing-data/

Conflict of interest

The authors declare that there is no conflict of interest.

References

Archer, G., Xun, W.W., Stuchbury, R., Nicholas, O. and Shelton, N. (2020) Are healthy cohorts real-world relevant? Comparing the National Child Development Study (NCDS) with the ONS Longitudinal Study (LS), *Longitudinal and Life Course Studies*, 11(3): 307–30. doi: 10.1332/175795920X15786630201754

Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J. (2011) Multiple imputation by chained equations: what is it and how does it work?, *International Journal of Methods in Psychiatric Research*, 20(1): 40–9. doi: 10.1002/mpr.329

Carpenter, J. and Kenward, M. (2012) *Multiple Imputation and Its Application*, Chichester: Wiley.

Harel, O., Mitchell, E.M., Perkins, N.J., Cole, S.R., Tchetgen Tchetgen, E.J., Sun, B. and Schisterman, E.F. (2018) Multiple imputation for incomplete data in epidemiologic studies, *American Journal of Epidemiology*, 187(3): 576–84. doi: 10.1093/aje/kwx349

Mostafa, T., Narayanan, M., Pongiglione, B., Dodgeon, B., Goodman, A., Silverwood, R.J. and Ploubidis, G.B. (2021) Missing at random assumption made more plausible: evidence from the 1958 British birth cohort, *Journal of Clinical Epidemiology*, 136: 44–54. doi: 10.1016/j.jclinepi.2021.02.019

White, I.R., Royston, P. and Wood, A.M. (2011) Multiple imputation using chained equations: Issues and guidance for practice, *Statistics in Medicine*, 30(4): 377–99. doi: 10.1002/sim.4067