Technical report on the enhancement of Millennium Cohort Study data with linked electronic health records; derivation of consent weights

Authors

Francesco Sera[1,2], Lucy J Griffiths[1], Carol Dezateux[1], Mario Cortina-Borja[1]

Affiliations

[1] Population, Policy and Practice Programme. UCL Great Ormond Street Institute of Child Health

[2] Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine

Scope

This document applies to the preparation of a Standard Operating Procedure (SOP) for the Wellcome Trust Data Linkage Project regarding the definition of consent weights for linkage to electronic health records between routinely collected data and data from the Millennium Cohort Study (MCS).

Date: 23 January 2018

Table of Contents

Funding Sources

The Millennium Cohort Study is funded by grants to the Centre for Longitudinal Studies at the Institute of Education from the Economic and Social Research Council and a consortium of government departments. The work described in this report relating to derivation of the consent weights was funded by the Wellcome Trust (grant title: “Using health record linkage in the UK Millennium Cohort Study to investigate childhood obesity, asthma and infections”; grant reference 087389/B/08/Z; Co-PIs: Professor Ronan Lyons and Professor Carol Dezateux). This work was undertaken at the Life Course Epidemiology and Biostatistics, UCL Great Ormond Street Institute of Child Health, London, which is supported by the National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London.

Acknowledgements

The authors are grateful to others involved in the work reported here: Colleagues at the Centre for Longitudinal Studies, UCL Institute of Education: Jon Johnson, Efrosini Setakis, Emla Fitzsimons; and those working at Swansea University within the Secure Anonymised Information Linkage (SAIL) system, which is part of the national e-health records research infrastructure for Wales: Ronan A Lyons, Amrita Bandyopadhyay, Karen Tingay and Ashley Akbari.

The co-operation of the participating families is gratefully acknowledged.

Ethical approval for the fourth survey of the Millennium Cohort Study was received from the Northern and Yorkshire Research Ethics Committee (07/MRE03/32).

Creative Commons License

Introduction

The Millennium Cohort Study (MCS) is a multidisciplinary survey of over 19,000 children born in the UK in 2000-01 who are followed over time. A disproportionately stratified clustered sampling design was used to over-represent children living in Wales, Scotland and Northern Ireland, disadvantaged areas and areas with high proportions of ethnic minority groups. The first survey took place when the children were aged around 9 months old,2 and subsequent surveys have taken place when the children were aged around 3 years, 5 years, 7 years, 11 years and 14 years old. The survey collects information from parents covering a range of domains including socio-economic circumstances, parenting, child's activities and behaviour, child and parental health, neighbourhood, relationships, childcare, and child's education and schooling. For further information on the MCS see: www.cls.ioe.ac.uk/mcs

At the fourth survey (~ 7 years) parents or carers were asked to give consent to linkage of information collected within MCS to their child's routine electronic heath records through to age 14 years. Further information on the consent procedure, including the consent form used and validation of the consents received has been reported by Shepherd (2013). This report describes the steps used to calculate the consent weights that should be used in analyses of linked MCS and electronic health data.

Consent Analysis

Productive families at MCS4

There were 13,857 productive families in the 4th sweep of the MCS (MCS4), for a total of 14,043 children.

Consent categories

There were 13,047 children consenting, 996 nonconsenting. Note that there are no missing values for this outcome variable. The outcome variable was named C and coded as 1 (Yes) and 0 (No).

Predictors of consent

Predictors of consent were initially identified among those used by (Ketende, 2010) to analyse nonresponse at MCS4, and by (Rich et al, 2013) to predict probability to participate to the MCS accelerometry study. Predictors were grouped as follows (Plewis, 2007):

1. UK country.
2. Socio-demographic variables: sex and ethnic group of the cohort member; if the cohort member is singleton or not; cohort member's sweep of MCS entry; main respondent's age at birth of the cohort member, main respondent's highest academic qualification, change of address between sweep 3 and 4.
3. Socio-economic variables: yearly net family income; housing tenure; socio-economic status (NS-SEC), type of accommodation and occupational status of the respondent.
4. Miscellaneous: whether there was a partner in the household, they had been interviewed; number of children in the household; household language; cohort member's longstanding illness; whether the mother had breast-fed the cohort child.

Statistical models

Predictors of non-consent to data linkage was obtained using logistic regression with *C* as a binary outcome variable (reference category: consent: *C*= 0). Odds ratios (ORs), adjusted odds ratios (aORs) and 95% confidence intervals (CI) were calculated for the predictors introduced above in both univariable and multivariable regression models. All models were fitted taking into account the complex sampling design used in the MCS (Plewis, 2007).

The final models were defined using the following model selection strategy. For the analysis of whole UK data, we first selected all variables which were significant (p < 0.05) in the univariable analyses. These variables were considered in the first multivariable model, and those which remained significant were included in the final multivariate model. In the country-specific analyses, all variables with an OR > 1.49 or OR < 0.67 (i.e. regression coefficient > 0.4 or < -0.4) were considered in the first multivariable model. The final model only included those variables which had an odds ratio within these bounds in the adjusted model.

Since there were incomplete cases in some predictors, a sensitivity analysis on the coefficients of the multivariable logistic regression models was performed by using a multiple imputation procedure. The standard error of the models' parameter estimates were obtained from a robust variance estimation procedure. The calculations were carried out on Stata version 13; we used the Stata routines **ice** (Carlin et al 2008) and **mi** (van Buuren 2007) to perform multiple imputation.

In all analyses, including the multiple imputation procedure, the MCS sampling design was taken into account using sampling weights adjusted for nonresponse to MCS4 (Plewis, 2007).

To develop country-specific weights, the final logistic regression multivariable models were used to estimate the probability of consent. In both single country and whole UK analysis multiple imputation was performed to take into account missing data on predictors. Country-specific sampling weights adjusted for nonresponse to MCS4 were used in the imputation procedures.

In all UK and country-specific analysis the predicted probability of consent was calculated as 1 minus the probability of non-consent estimated from multivariable logistic model. These calculated probabilities were multiplied by the longitudinal MCS4 weights giving the longitudinal weights for the data linkage MCS4 study. The weights were scaled to have as a sum the number of consenting MCS4 children (e.g. 13,047 for the all UK analysis). These consent weights should be used in the analyses including data from the all UK MCS cohort linked to other databases, e.g. Hospital Episode Statistics.

Results

Whole UK analysis

The results in Table 1 concern the distribution of the 14,043 MCS4 children consenting (13,047) and not consenting (996), and show unadjusted and adjusted odds ratios (OR and aOR) and *p*-values obtained from the final fitted multivariable logistic regression models following multiple imputation. The following groups showed a significantly lower probability of consent:

1. Children of Pakistani/Bangladeshi and Black/Black British ethnicity had over a two-fold increase in their odds of consenting.
2. Children living in households defined as flat/maisonette/studio/room/bedsit had 45% increased odds of consenting.
3. Children living in households whose yearly net income is > £31,200 (the fourth quintile of the household income distribution) had around 40% higher odds of consenting.

4. Children living in households where the main respondent was only one of the participant's parents or a person responding by proxy had an almost three-fold increase in their odds of consenting.
5. Children living in households with one child had 33% higher odds of consenting compared to those in households with two or three children.

Country-specific analysis

Table 2 presents odds ratios and 95% CIs estimated from the final multivariable logistic regression models obtained following the selection strategy outlined in the statistical methods section, and following multiple imputation separately for England, Wales, Scotland and Northern Ireland. Note that only predictors with significant regression coefficients ($p <$ 0.05) are shown in the table. In this case the weights were defined on much smaller sets of predictors than those in Table 1; this is as expected, given the degree of between-country heterogeneity. The following groups defined the weights' models by country:

1. In England, the same variables defining the models in the whole UK estimation were included: ethnicity, type of accommodation, household income, living in households where the main respondent was only one of the participant's parents or a person responding by proxy, and number of children in the household played the same role as that in the models for data for the whole of the UK, as detailed above.
2. In Wales the weights were defined only in terms of living in households where the main respondent was only one of the participant's parents or a person responding by proxy increased the odds of consenting by more than three times.
3. In Scotland, comparing with households where both parents responded, the odds of consenting were 75% higher among single-parent households, and over three times higher where the main respondent was only one of the participant's parents or a person responding by proxy. In addition, if the household did not have a stable address, decreased 45% the odds of consenting.
4. In Northern Ireland comparing with households where both parents responded, the odds of consenting were over 4 times higher among households where the main respondent was only one of the participant's parents or a person responding by proxy. Age of the main respondent was also determinant with those under 19 years having odds of consent 90% smaller compared with respondents over 30 years. Type of

accommodation was also in the model, with households defined as flat/maisonette/studio/room/bedsit having a three-fold increase of the odds of consenting, much larger than the increase of 40% observed for England. Households without a stable address had an increase of 3.3 times in the odds of consenting, in the opposite direction of the decrement observed in Scotland.

The minimum, maximum, mean and standard deviation of sampling and non-response adjusted weights according MCS sampling stratum for UK, and for separate countries are shown in Tables 3 and 4. These summary statistics are broken down by the three types of district strata (advantaged, disadvantaged and ethnic) defined in the MCS sampling design.

The weights were stored in the variables **dovwt2_Linkage** (whole UK analysis) and **dovwt1_Linkage** (single country analyses)

The minimum, maximum, mean and standard deviation of sampling and non-response adjusted weights according MCS sampling stratum are represented on Table 4.

Conclusions

The consent weights described in this document should be used in analyses of linked MCS and health data. Whole UK, or country-specific weights are available and should be used accordingly. Both sets of weights are available at the UK data service.

References

Carlin, J. B., Galati, J. C., Royston, P. (2008). A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal* 8(1):49-67.

Ketende SC. (2010) *The Millennium Cohort Study: Technical Report on Response. 3rd ed.*Centre for Longitudinal Studies, London.

Plewis I. (2007) *The Millennium Cohort Study: technical report on sampling. Technical Report 4th edition*. Centre for Longitudinal Studies, London.

Rich C, Cortina-Borja M, Dezateux C, Geraci M, Sera F, Calderwood L, Joshi H, Griffiths LJ (2013)Predictors of non-response in a UK-wide cohort study of children's accelerometer-determined physical activity using postal methods. *BMJ Open*. 2013 Mar 1; 3 (3).

Shepherd P. (2013) *Millennium Cohort Study: Consent to linkage to child health data.* Centre for Longitudinal Studies, London.

van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16: 219–242.

Tables

Table 1. Distribution of the 14,043 MCS4 children (consenters and non-consenters) according to predictors of consent with unadjusted and adjusted odds ratios estimated with logistic regression after multiple imputation procedure with robust variance estimation of the parameters. The models refer to the whole of the UK sample.

			Consent Health Linkage		**Univariable**		**Multivariable**	
Variable		**Total**	**Yes** **(*n* = 13047)**	**No** **(*n* = 996)**				
			%	**%**	**OR**	***p*-value**	**OR**	***p*-value**
Main respondent's age (years) at the birth of the cohort member	**14-19**	1011	94.1	5.9	0.87	0.345		
	20-29	6219	92.5	7.5	1.12	0.213		
	30+	6813	93.2	6.8				
Main respondent's education	**Degree**	2834	92.2	7.8	1.32	0.039		
	Diploma in higher education	1557	93.1	6.9	1.15	0.365		
	A/As/S levels	1306	93.3	6.7	1.12	0.344		
	GCSE grades A-G	5718	94.0	6.0				
	Other academic qualifications	386	90.9	9.1	1.56	0.037		
	None of the above	2199	91.4	8.6	1.47	0.002		
When Joined MCS	**Sweep 1**	13543	92.9	7.1				
	Sweep 2	500	93.4	6.6	0.92	0.706		
Stable address	**Yes**	12664	92.9	7.1				
	No	1378	93.2	6.8	0.96	0.727		
Cohort member breast-fed	**Yes**	9511	93.7	6.4				
	No	4470	92.8	7.2	0.87	0.105		
Country	**England**	8955	93.1	6.9				
	Wales	2039	94.2	5.9	0.84	0.268		

	Scotland	1654	91.2	8.8	1.31	0.064		
	Northern Ireland	1395	92.3	7.7	1.13	0.425		
Cohort child's ethnic group	**White**	11691	93.8	6.2				
	Mixed	379	91.9	8.1	1.33	0.272	1.23	0.418
	Indian	343	90.9	9.1	1.51	0.111	1.60	0.088
	Pakistani/Bangladeshi	880	86.7	13.3	2.31	0.000	2.27	0.000
	Black/Black British	423	84.8	15.2	2.71	0.000	2.34	0.000
	Other	223	91.4	8.6	1.43	0.286	1.28	0.438
Cohort child's sex	**Male**	7110	92.6	7.4				
	Female	6933	93.3	6.7	0.90	0.137		
Whether main respondent is in work or not	**Is in work**	8744	93.4	6.6				
	Not in work or leave	5299	92.3	7.7	1.18	0.077		
Housing tenure	**Own/Mortgage**	9257	93.3	6.7				
	Rent	4366	93.5	6.5	0.97	0.748		
	Other	277	92.0	8.0	1.21	0.475		
Type of accommodation	**House or bungalow**	12772	93.4	6.6				
	Flat, maisonette, studio, room, bedsit,	1250	88.8	11.2	1.79	0.000	1.45	0.003
Household annual income	**1040-10400**	1755	91.8	8.3	1.27	0.062	1.19	0.197
	10400-20800	3897	93.4	6.6				
	20800-31200	3312	93.6	6.4	0.97	0.781	1.14	0.202
	31200-52000	3464	92.9	7.1	1.08	0.575	1.40	0.017
	52000+	1595	92.7	7.3	1.11	0.547	1.46	0.027
Parents response summary	**Single Parent**	2940	93	7	1.28	0.040	1.22	0.158
	Both parents	9289	94.4	5.6				
	One of the two or no parents	1814	85.1	14.9	2.97	0.000	2.82	<0.001
NS-SEC	**Managerial and professional occupations**	4058	92.3	7.7	1.15	0.210		
	Intermediate occupations	2429	94.2	5.8	0.86	0.192		

	Small employers and own account workers	911	93.5	6.5	0.96	0.835		
	Lower supervisory and technical occupations	626	94.3	5.7	0.84	0.409		
	Semi-routine and routine occupations	4672	93.3	6.7				
	Not at work or long term unemployed	1206	90.3	9.7	1.49	0.008		
Number of children in the household	**1**	1771	90.9	9.1	1.36	0.005	1.33	0.016
	2-3	10141	93.2	6.9				
	>=4	2131	93.9	6.1	0.89	0.396	1.30	0.045
Household language	**English only**	12115	93.6	6.5				
	Mostly English	1350	88.1	11.9	1.96	0.000		
	Mostly other	578	87	13.0	2.17	0.001		
Longstanding illness	**No**	11348	93.3	6.7				
	Yes	2613	93.1	6.9	1.03	0.764		

*Due to small sample size it was not possible to estimate the standard error for this category

Table 2. UK country specific odds ratios estimated with multivariable logistic regression after multiple imputation procedure with robust variance estimation of the parameter.

		England		Wales		Scotland		Northern Ireland	
		OR	***p*- value**	**OR**	***p*- value**	**OR**	***p*- value**	**OR**	***p*- value**
Main respondent's age at the birth of the cohort member	**14-19**							0.10	0.003
	20-29							0.86	0.463
	30+								
Stable address	**Yes**								
	No					0.45	0.026	3.31	0.001
Cohort child's ethnic group	**White**								
	Mixed	1.33	0.275						
	Indian	1.60	0.094						
	Pakistani/Bangladeshi	2.54	0.000						
	Black/Black British	2.60	0.000						
	Other	1.55	0.207						
Type of accommodation	**House or bungalow**								
	Flat, maisonette, studio, room, bedsit,	1.39	0.022					2.99	0.034
Household annual income	**1040-10400**	1.26	0.151						
	10400-20800								
	20800-31200	1.19	0.167						
	31200-52000	1.49	0.023						
	52000+	1.60	0.021						
Parents response summary	**Single Parent**	1.20	0.294	1.21	0.609	1.75	0.032	1.35	0.319
	Both parents								
	One of the two or no parents	2.63	0.000	3.38	0.000	3.14	0.000	4.39	0.000
No. of children in the household	**1**	1.42	0.012						
	2-3								
	>=4	0.75	0.071						

Table 3. MCS1 and MCS4 minimum, maximum, mean and standard deviation (SD) sampling and non-response adjusted weight estimates for analyses of the whole of UK sample.

		Sampling weights	**Overall weights wave 4 (dovwt2)**				**Overall weights wave 4 Record Linkage study (dovwt2_Linkage)**			
	n	**weight2**	**min**	**max**	**Mean**	**SD**	**min**	**max**	**Mean**	**SD**
England - advantaged	3593	2.00	1.24	5.45	1.71	0.40	1.24	5.41	1.70	0.41
England - disadvantaged	3247	1.09	0.72	4.99	1.17	0.40	0.70	6.71	1.17	0.43
England - ethnic	1480	0.37	0.26	2.70	0.51	0.20	0.25	2.70	0.53	0.22
Wales - advantaged	595	0.62	0.41	1.78	0.56	0.15	0.40	1.87	0.55	0.15
Wales - disadvantaged	1328	0.23	0.15	0.70	0.24	0.07	0.14	0.69	0.24	0.07
Scotland - advantaged	787	0.93	0.36	3.17	0.89	0.37	0.35	3.05	0.89	0.37
Scotland - disadvantaged	727	0.57	0.22	1.97	0.67	0.31	0.22	2.14	0.67	0.32
Northern Ireland - advantaged	507	0.47	0.19	1.75	0.51	0.20	0.18	1.85	0.51	0.21
Northern Ireland - disadvantaged	783	0.25	0.10	1.13	0.32	0.14	0.10	1.11	0.32	0.14

Notes:

Weight2: whole UK sampling weight

dovwt2: the longitudinal weight at sweep 4 which is a product of sweep 3 overall weight (**covwt2**) and non-response weight at sweep 4

dovwt2_Linkage: the longitudinal weight for consent study at sweep 4 which is a product of sweep 4 overall weight (**dovwt2**) and consent study non-response weight

Table 4. MCS1 and MCS4 minimum, maximum, mean and standard deviation (SD) sampling and non-response adjusted weight estimates for UK country-specific analysis.

		Sampling weights	**Overall weights wave 4 (dovwt1)**				**Overall weights wave 4 Record Linkage study (dovwt1_Linkage)**			
	n	(**weight1**)	**min**	**max**	**mean**	**SD**	**min**	**max**	**mean**	**SD**
England - advantaged	3593	1.32	0.98	4.29	1.34	0.31	0.98	4.24	1.33	0.32
England - disadvantaged	3247	0.71	0.56	3.88	0.91	0.31	0.55	5.20	0.91	0.33
England - ethnic	1480	0.24	0.20	2.09	0.39	0.16	0.19	2.08	0.41	0.17
Wales - advantaged	595	1.77	1.20	5.23	1.64	0.44	1.19	5.58	1.64	0.46
Wales - disadvantaged	1328	0.65	0.43	2.06	0.71	0.20	0.42	2.06	0.71	0.20
Scotland - advantaged	787	1.23	0.46	4.04	1.14	0.47	0.45	3.91	1.14	0.48
Scotland - disadvantaged	727	0.75	0.28	2.49	0.84	0.39	0.28	2.72	0.85	0.40
Northern Ireland - advantaged	507	1.41	4.39	1.28	0.51	0.46	0.46	4.69	1.28	0.52
Northern Ireland - disadvantaged	783	0.76	0.26	2.85	0.81	0.35	0.26	2.84	0.82	0.36

Notes:

weight1: sampling weight

dovwt1: the longitudinal weight at sweep 4 which is a product of sweep 3 overall weight (**covwt1**) and non-response weight at sweep 4

dovwt1_Linkage: the longitudinal weight for consent study at sweep 4 which is a product of sweep 4 overall weight (**dovwt1**) and consent study non-response weight.