Knitting studies together: combining longitudinal trajectories from different cohorts to examine the life-course trajectory

Hughes RA[1,2], Tilling K[1,2], Lawlor DA[1,2]

[1]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[2]MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom

Correspondence: *Dr Rachael Hughes, MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, UK, BS8 2PS. E-mail: rachael.hughes@bristol.ac.uk*

BACKGROUND

- Longitudinal data analysis is necessary to reveal changes within the same individual as they age. However, few studies are able to capture multiple decades across the life-course.
- We investigate the challenges in combining data from cohorts with repeated measurements that cover different and overlapping periods of life. Our illustrative example examines the effects of gender, ethnicity and maternal education on weight trajectories.

METHODS

- We analysed data from five prospective cohorts: Avon Longitudinal Study of Parents And Children[1,2] (ALSPAC), Born in Bradford[3] (BiB), Barry Caerphilly Growth (BCG) study[4,5], Christ Hospital Study[6] (CHS), and the PRomotion Of Breastfeeding Intervention Trial[7,8] (PROBIT). Children from multiple births were excluded because their growth patterns differ considerably from singletons. Data were harmonized to ensure the definitions and classifications of covariates were inferentially equivalent.

Statistical methods

We considered fractional polynomials and natural splines (also known as restricted cubic splines) to describe the relationship between weight and age (i.e., the growth trajectory). Both offer a greater range of curve shapes than linear or quadratic polynomials. We investigated fractional polynomials of one and two degrees with powers {–2, –1, –0.5, 0, 0.5, 1, 2, 3}, including models with repeated powers, and natural splines with 3 to 7 knots, where the knots were placed at equally spaced percentiles of age[9]. The growth trajectory was fitted as a random effects model, with the intercept and fractional polynomial or natural spline terms random at the child-level, allowing growth trajectories to vary between children. Cohort was included as a fixed effect, with interactions between cohort and the trajectory terms considered. Additionally, we allowed the observation-level variance to vary with age by considering models with age random at the observation level, and splitting the age range into periods with the observation level variance constant within a period but allowed to differ across periods. We included covariates gender, ethnicity and maternal education as fixed effects. For each covariate, interactions between the covariate and trajectory terms were considered.

The best fitting model was selected by comparing the Bayesian Deviance Criterion (BIC) of different models and the Mean Squared Error of the predicted values (i.e., mean of the squared differences between the observed and predicted values). Due to the large amount of data we conducted model selection on a random sample of the combined data, where the proportion of children from each cohort was the same in the random and combined samples. The alternative strategy of conducting model selection separately within each cohort was infeasible due to varying age-periods across the cohorts. After model selection, all remaining analyses were conducted on all of the combined data.

Missing child-level data were multiply imputed (100 times) at the child-level assuming data were missing at random. The imputation model included cohort, child-level variables (gender, ethnicity, paternal SES and maternal education) and cluster means of the observation-level variables (e.g., within-child mean of weight, age spline terms) to account for the multilevel structure of the data[10], plus cluster count (within-child number of measures) and its interaction with the cluster means to account for varying cluster sizes[10]. The interaction between age and maternal education was incorporated by including cluster means of the interaction between weight and the age spline terms[11]. The 100 sets of results from analyzing the imputed datasets were combined into a single inference using Rubin's rules.

RESULTS

- The cohorts contributed data from birth to age 20 years. The combined sample size was 47,205 children (53% boys) with 542,781 weight measurements.
- The cohorts were based in four distinct regions of the United Kingdom and the Republic of Belarus, with most of the data collected from the early-1990s onwards (table 1). The population represented by BiB has high levels of socioeconomic deprivation, in contrast to cohort CHS which was based in a private male boarding school. Most children were white, and approximately half were boys.
- Maternal educational (highest attainment) was not recorded by cohorts BCG and CHS, and was only partially observed by cohorts ALSPAC and BiB. Also, across all cohorts, paternal Socio-Economic Status (SES) was missing for 14% of the children.

TABLE 1: BASELINE CHARACTERISTICS OF THE COHORTS

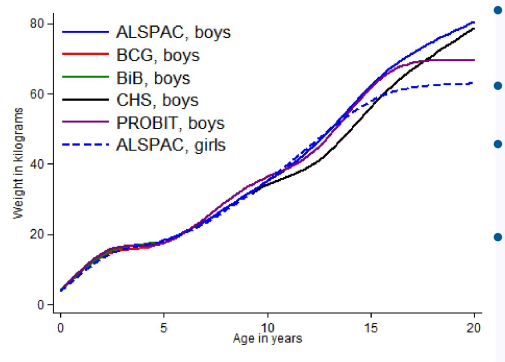
	ALSPAC	BCG	BiB	CHS	PROBIT
Region	South West England	South East Wales	Centre North of England	South East England	Republic of Belarus
Calendar period	1990 to 2012	1972 to 1979	2007 to 2015	1936 to 1964	1996 to 2016
Age range	0 to 20 years	0 to 5 years	0 to 7 years	9 to 18 years	0 to 19 years
No. subjects	14, 216	951	13,445	1,547	17,046
Gender					
Male	$51%	54%	52%	100%	52%
Ethnicity					
White	79%	100%	36%	100%	100%
South Asian	0%	0%	44%	0%	0%
Other	21%	0%	20%	0%	0%
Maternal education					
Left school at 15 or 16	55%	0%	43%	0%	4%
Left school at 17 or 18	19%	0%	12%	0%	82%
Degree	11%	0%	21%	0%	14%
Missing	15%	100%	24%	100%	0%
Paternal SES#					
Class I or II	22%	17%	13%	55%	11%
Class III	36%	58%	19%	21%	60%
Class IV, V or other	21%	23%	44%	3%	25%
Missing	21%	2%	24%	21%	4%

$ Excludes two subjects with missing values for gender. # Socio-Economic Status (SES) based on father's occupation.

RESULTS (continued)

- The harmonisation versions of paternal SES and maternal education coarsened these variables to 3 categories, potentially losing information. Due to the known lower birth weights and growth rates among South Asians compared to white and other ethnicities, variable ethnicity was harmonised to white, South Asian and other, even though only BiB recorded children from a South Asian heritage.
- The best fitting model was a natural spline with 7 knots at 0.25, 2.5, 4.5, 9.25, 12.25, 15.5 and 18.75 years. All two-way interactions between the covariates and the trajectory terms were included. We excluded all three-way and higher interactions.
- Minimal differences in population growth trajectories between categories of ethnicity and maternal education.

FIGURE 1: ESTIMATED POPULATION GROWTH TRAJECTORIES ACCORDING TO COHORT AND GENDER, WHERE REMAINING VARIABLES WERE SET TO REFERENCE VALUES (WHITE AND MOTHER LEFT SCHOOL AT 15 OR 16)



- Model assumes differences between boys and girls were the same in each cohort.
- Gender differences became apparent after age 14 (figure 1).
- The plateauing effect after age 15 for the PROBIT cohort could be due to its limited number of measures between 15-20 years.
- Lower growth trajectory for CHS, compared to ALSPAC and PROBIT, could be due to a changes in diet between the mid-1900s and the late 1990s/early 2000s.

TABLE 2: COMPLETE CASE ANALYSIS AND MULTIPLE IMPUTATION RESULTS FOR THE INTERACTION BETWEEN MATERNAL EDUCATION AND THE INTERCEPT AND SPLINES

	Complete case analysis (39,374 children with 412,284 measures)		Multiple imputation (47,203 children with 542,779 measures)	
	Estimate (SE*)	[LCI£, UCI§]	Estimate (SE)	[LCI, UCI]
Left school at 17 or 18: intercept	0.080 (0.0087)	[0.063, 0.097]	0.074 (0.0088)	[0.057, 0.091]
age spline 1	-0.085 (0.017)	[-0.18, -0.052]	-0.065 (0.016)	[-0.097, -0.033]
age spline 2	3.6 (0.45)	[2.8, 4.5]	3.0 (0.45)	[2.1, 3.9]
age spline 3	-11 (1.4)	[-14, -8.6]	-9.3 (1.5)	[-12, -6.4]
age spline 4	9.7 (1.4)	[7.0, 12]	7.5 (1.6)	[4.4, 11]
age spline 5	-2.6 (1.5)	[-5.5, 0.38]	0.12 (1.8)	[-3.5, 3.7]
age spline 6	0.46 (2.8)	[-5.0, 5.9]	-4.0 (3.2)	[-10, 2.2]
Degree: intercept	0.11 (0.0091)	[0.095, 0.13]	0.10 (0.0091)	[0.085, 0.12]
age spline 1	-0.041 (0.018)	[-0.075, -0.0060]	-0.018 (0.018)	[-0.053, 0.016]
age spline 2	2.98 (0.48)	[2.0, 3.9]	2.2 (0.49)	[1.2, 3.1]
age spline 3	-8.4 (1.5)	[-11, -5.4]	-5.7 (1.6)	[-9.0, -2.5]
age spline 4	5.9 (1.5)	[2.9, 8.9]	3.1 (1.6)	[-0.44, 6.6]
age spline 5	1.0 (1.7)	[-2.3, 4.4]	4.9 (2.1)	[0.72, 9.1]
age spline 6	-3.9 (3.2)	[-10, 2.4]	-11 (3.8)	[-19, -3.9]

*SE: Standard Error; £LCI: Lower 95% Confidence Interval; §UCI: Upper 95% Confidence Interval

- Complete case analysis (CCA) and multiple imputation (MI) make different assumptions about the missing data. CCA allows missingness to depend on the unobserved maternal education values but not paternal SES since it is not a covariate of our model of interest. Whilst MI assumes missingness only depends on observed data but allows missingness to depend on paternal SES which is an auxiliary variable of the imputation model.
- MI estimates tended to be closer to the null than CCA estimates. Also MI had slightly larger standard errors than CCA.
- Both analyses led to the same conclusions about the effect of maternal education on childhood growth.

CONCLUSIONS

- This approach enables modelling of trajectories over wide age ranges and sharing of information across studies. However, this approach may be infeasible when studies do not measure the same repeated outcome.
- Multiple imputation can be used to impute variables that were partially observed by some cohorts and not measured at all by other cohorts. However, imputing large amounts of data could introduce bias and inflate standard errors if there are insufficient predictors of the missingness process and missing data.
- Including more variables in the imputation model can reduce bias and improve precision. However, it may be difficult to identify variables measured by all cohorts that can be appropriately harmonized for analysis.
- With modern computing power and software, these complex analyses on a very large dataset are feasible.

REFERENCES AND ACKNOWLEDGMENTS

(1) Fraser A et al. Int J Epidemiol 2013;42:97-110. (2) Boyd A et al. Int J Epidemiol 2013;42:111-27. (3) Wright J et al. Int J Epidemiol 2013;42:978-91. (4) Elwood PC et al. Arch Dis Child 1981;56:831-35. (5) McCarthy A et al. Am J Clin Nutr 2007;86:907-13. (6) Sandhu J et al. Int J Obes 2006;30:14-22. (7) Kramer MS et al MS. JAMA 2001;285:413-20. (8) Patel R et al. Int J Epidemiol 2014;43:679-90. (9) Harrell FE, Jr. 2001 New York: Springer. (10) Grund S et al. J Educ Behav Stat 2018;43:316-51. (11) Tilling K et al. J Clin Epidemiol 2016;80:107-15. (12) Leckie G et al. J Stat Soft 2013;52:1-40.

We thank the following for their permission to analyse data from their cohorts: the ALSPAC executive committee, the BiB executive committee and BiB analysis co-ordination group, Professor Yoav Ben-Shlomo (University of Bristol) for BCG and CHS, and Professor Michael Kramer (McGill University) and Professor Richard Martin (University of Bristol) for PROBIT. We are extremely grateful to all families that took part in these cohorts, the cohort teams and clinicians who helped with recruitment and data collection. No individual participant data were reported by this study. .

This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 733206