Combining longitudinal data from different cohorts to examine the life-course trajectory

Rachael A. Hughes¹,², Kate Tilling¹,², Deborah A. Lawlor¹,²,³

¹MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, UK,
²Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK
³Bristol NIHR Biomedical Research Centre, Bristol, UK.

Longitudinal data are necessary to reveal changes within the same individual as they age. However, rarely will a single cohort capture data throughout the lifespan. We describe in detail the steps needed to develop life-course trajectories from cohorts that cover different and overlapping periods of life. Such independent studies are likely from heterogeneous populations which raises several challenges including: data harmonisation (deriving new harmonised variables from differently measured variables by identifying common elements across all studies); systematically missing data (variables not measured are missing for all participants of a cohort); and model selection with differing age ranges and measurement schedules. We illustrate how to overcome these challenges using an example which examines the effects of parental education, sex, and ethnicity on weight trajectories. Data were from five prospective cohorts (Belarus and four UK regions), spanning from birth to early adulthood during differing calendar periods. Key strengths of our approach include modelling trajectories over wide age ranges, sharing of information across studies and direct comparison of the same parts of the life-course in different geographical regions and time periods. We also introduce a novel approach of imputing individual-level covariates of a multilevel model with a nonlinear growth trajectory and interactions.

Keywords: ALSPAC, Barry Caerphilly Growth study, Born in Bradford, Christ Hospital study, life-course, mixed-effects models, PROBIT, repeated measures.
Life-course epidemiology aims to elucidate biological, behavioural, and psychosocial processes that operate across an individual’s life-course to influence the development of disease risk [1]. It requires repeatedly assessed data on risk factors and trajectories that reflect the underlying propensity of disease risk from infancy to adulthood [1]. Trajectories of underlying disease propensity are necessary to establish ages of ‘peak’ health for different diseases and the extent to which intervening during periods of developmental origins or age-related decline are likely to maximise population health. However, rarely will a single cohort have data on a given risk factor or outcome throughout the lifespan.

An alternative to examining life-course trajectories in a single cohort is to model longitudinal data from multiple cohorts that cover different and overlapping periods of life. Examples include models of longitudinal blood pressure data from eight UK cohorts from ages 7 to 80 [2], alcohol consumption from age 15 years to over 90 years from nine cohort studies [3] and height and weight trajectories from birth to 18 years using sparse longitudinal data from 16 studies [4]. Whilst these studies demonstrate the potential of obtaining life-course trajectories by combining data from several cohorts that cover different periods of the life-course, they do not describe the steps required to overcome the challenges in conducting this type of analysis.

This paper investigates the challenges in combining data from independent cohorts with repeated measurements that cover different and overlapping periods of life. We describe in detail the steps needed to model life-course trajectories from several cohorts (summarised in box 1), including our code in statistical software Stata.

ILLUSTRATIVE EXAMPLE
Our example examines the associations of sex, ethnicity, and parental education on weight trajectories from birth to 20 years. We analysed data from five prospective cohorts: the Avon Longitudinal Study of Parents and Children (ALSPAC) [5, 6], Barry Caerphilly Growth (BCG) study [7, 8], Born in Bradford (BiB) [9], Christ’s Hospital School (CHS) [10], and the PRomotion of Breastfeeding Intervention Trial (PROBIT) [11, 12]. PROBIT is from Belarus and the remaining cohorts are from the UK. Collectively they cover different periods of birth from the 1930s to 2010. Children from multiple births were excluded because their growth patterns differ considerably from singletons. Parental education is a composite variable that takes on all possible joint values of maternal highest educational attainment and paternal occupation (recorded around the time of birth except for CHS). Data were harmonised across the cohorts.

We give a brief description of each cohort study (further details in the Appendix). Table 1 summarises the cohorts’ characteristics.

**Cohort ALSPAC**

All pregnant women resident in a defined area in the South West of England, with an expected date of delivery between 1st April 1991 and 31st December 1992, were invited to take part in ALSPAC [5, 6]. The children of these pregnancies have been followed up since birth. Weight measurements were recorded at multiple timepoints between birth and late adolescence (e.g., at birth, 6 weeks, 10, 21 and 48 months, annual measurements between 7 to 11 years, and at target ages 12, 13, 15 and 17 years) [13]. These measurements were from several sources: medical records, research clinics and parent-reports.

**Cohort BCG**
BCG is a follow-up of a dietary intervention randomised controlled trial of pregnant women and their offspring [8], which recruited pregnant women resident in two towns in South Wales between 1972 and 1974 [7]. Birth weights were abstracted from hospital records, and weight measurements were recorded by study nurses who visited at 10 days, 6 weeks, 3, 6, 9 and 12 months, and thereafter at 6-monthly intervals resulting in a total of 14 measurements by age 5 years [8].

**Cohort BiB**

All pregnant women booked for delivery at the Bradford Royal Infirmary between March 2007 and November 2010, who attended the oral glucose tolerance test clinic (offered to all women at 26-28 weeks gestation), were invited to take part in the study [9]. The children of these pregnancies have been followed up since birth. Weight measurements were recorded at multiple timepoints between birth and mid-childhood (e.g., birth and taken on average at 2 and 6 weeks, 8 months, 4 and 6 years) [13]. These measurements were from several sources: maternity records, child health records, primary care records, the national child measurement programme (NCMP) and researcher assessments.

**Cohort CHS**

Regular measurements were recorded on boys, aged 9 to 18 years, who attended the (all boys) Christ’s Hospital school in West Sussex between 1936 and 1969 [10]. The students were measured three times a school term (i.e. nine measurements per year) by the School medical officer [10].

**Cohort PROBIT**
Thirty-one maternity hospitals and associated polyclinics (outpatient clinical for routine healthcare) in the Republic of Belarus participated in this randomised controlled trial. Mother-infant pairs were recruited during their postpartum hospital stay between June 1996 and December 1997 [11, 12]. Birth weight was abstracted from hospital records and weight was measured at scheduled study visits at 1, 2, 3, 6, 9 and 12 months, and at 6.5, 11.5 and 16 years [14]. Also, weight measurements were abstracted from primary care records.

CHALLENGE: DATA HARMONISATION

Data harmonisation is required when pooling heterogeneous data from different studies [15]. For comparable but differently measured variables, such as ethnicity, data harmonisation derives new ‘harmonised’ variables by identifying common elements across all studies.

Continuous measurements and sex

The continuous measurements, age and weight, only require conversion to the same units. All four birth cohorts recorded assigned sex at birth. CHS is a cohort from an all-boys school.

Child ethnicity

For cohorts BCG, CHS and PROBIT, who did not record ethnicity, we assumed all children were of White European origin because there were very few people from minority ethnic backgrounds in these populations during the cohort recruitment periods. We harmonised the BiB variable to categories “South Asian” and “White European” (representing 50% and 39% of the cohort, respectively), and to “other” for the 8% of Black, mixed race and other ethnicities (Appendix-table 1). And, we harmonised the ALSPAC variable to categories “White European” (79% of the cohort) and “other” (for the 4% from a Black African/Caribbean, South Asian, Chinese, mixed and ethnic backgrounds [16]).
Maternal education

Since qualifications can differ between countries and, between calendar periods within a
country (e.g., in 1987 the UK Ordinary-Level qualification was replaced by the General
Certificate of Secondary Education), we defined the harmonised variable by the number of
years the mother had spent in education (Appendix-table 2). The two older cohorts, BCG and
CHS, did not measure maternal education and so all of their values were set to missing.

Paternal occupation

BCG and CHS classified paternal Occupational Social Class based on the UK registrar
general classification [17] (Appendix-table 3). For ALSPAC and BiB, paternal occupation
was categorised using the Office for National Statistics’ Standard Occupational Classification
(of 1990 and 2000, respectively). Using [18] as a guide, we mapped the ALSPAC and BiB
Standard Occupational Classifications to three occupation classes compatible with those of
BCG and CHS (professional or managerial; intermediate; routine or unskilled). The
occupation categories of PROBIT cut across social classes (e.g., “service worker” includes
occupations from professional to routine, and “manual worker” includes skilled and unskilled
manual occupations). So, we used paternal occupation along with paternal highest
educational attainment to classify PROBIT fathers to “professional or managerial”,
“intermediate” or “routine or unskilled”.

METHODS

Challenge: accounting for the dependence structure of the data

The data has a 3-level structure with repeated measurements nested within an individual and
individuals nested within a cohort.
The data has a nested 3-level structure (repeated measurements nested within an individual and individuals nested within a cohort). There are two main approaches for modelling multilevel data: marginal models (e.g., generalized estimating equations models) and mixed effects models (also known as random effects, multilevel and hierarchical models) [19, 20]. We shall focus on the linear mixed effects model (hereafter referred to as a multilevel model) which can be used to make inferences about changes in the population mean response and to examine the data’s dependence structure (e.g., comparison of within-individual to between-individual variability).

Briefly, a multilevel model consists of “fixed effects” and “random effects” [21]. The fixed effects describe the average relationship between the repeated measurements and time, which we shall call the “average growth trajectory”. The random effects are defined at each level of the data and describe the multiple sources of random variability in the data (e.g., random variability between cohorts, between individuals of the same cohort, and within an individual).

Restricted maximum likelihood estimation is recommended because maximum likelihood estimation of the variance parameters (e.g., variance of a random effect) are biased downwards [21].

In our example, an important consideration was the small number of units at level-3 (i.e., 5 cohorts). Although, there is no clear rule of thumb regarding sample size requirements [22], a simulation study investigating multilevel modelling of units with 1000 observations each showed that at least 25 units was required for reliable inference about the variance parameters.
Therefore, we decided to fit a 2-level model (repeated measurements at level-1 and children at level-2) with cohort included as a categorical variable in the fixed effects with interactions between the cohort variable and the trajectory terms to allow each cohort to have its own average growth trajectory.

**Modelling the nonlinear growth trajectory**

Most biological growth processes show nonlinear changes over time [13]. Two common approaches to modelling nonlinear growth are fractional polynomials [23] and restricted cubic splines (also known as natural splines) [24].

A natural spline divides the growth trajectory into distinct segments where adjacent segments are joined at knot points. A separate curve is fitted to each segment, where the first and last segments are restricted to be linear, and cubic polynomials for the interior segments.

A fractional polynomial models the entire trajectory using a single, special type of polynomial that can include logarithms (e.g., $time^0 = \ln(time)$), minus powers (e.g., $time^{-2} = 1/time^2$), non-integer powers (e.g., $time^{0.5} = \sqrt{time}$) and repeated powers (e.g., $time^2, time^4$). The degree of the model indicates the number of fractional polynomial terms allowed (e.g., $time^2$ is a one-degree fractional polynomial, and $time^2, time^3$ is a two-degree fractional polynomial).

In our example, we considered natural splines with 3 to 7 knots, where the knots were placed at percentiles (of age) recommended by Harrell [24], and fractional polynomials of one and two degrees with powers -2, -1, 0, 0.5, 1, 2, 3.
Modelling the covariance structure

The covariance structure of the multilevel model accounts for the dependency within the data and is modelled via the random effects. Misspecification of the covariance structure can lead to incorrect standard errors for the fixed effects and invalid estimates of the random effects’ variance components [21].

At each level, the random effects are assumed to be normally distributed with zero means and an estimated variance-covariance matrix, where the diagonal elements are the variances of the random effects and the off-diagonal elements are pairwise covariances between the random effects. Many software implementations allow the analyst to place restrictions on the variance-covariance matrix (e.g., restrict a covariance to be 0). Typically, the random effects at each level are assumed to be mutually independent of the random effects at all other levels.

Including individual-level random effects for the intercept and slope term(s) allows the growth trajectories (i.e., starting positions and rate of change) to vary between individuals. Other variables, such as sex, can also be included as individual-level random effects.

The measurement-level random effects describe how an individual’s measurements vary about his/her growth trajectory. The simplest structure is a single random intercept which assumes the measurement-level variation is constant throughout. Alternatively, the model can allow the measurement-level variation to depend on age (known as complex level-1 variation [13]).
In our example, all models included individual-level random effects for the intercept and all trajectory terms. We did not put any constraints on the individual-level variance-covariance matrix (i.e., distinct variances and covariances were allowed).

Assuming a constant measurement-level variance was implausible for our data due to the wide age span (i.e., weights at age 18 years were more variable (e.g., 45-75 kg) than birth weights (e.g., 2.0-4.5 kg)). Therefore, we allowed for complex-1 variation, comparing two approaches: (1) included an intercept and linear term for age as measurement-level random effects. (2) Divided the age range into distinct segments and specified a measurement-level random intercept for each segment, where measurement-level variance could vary across the segments but was assumed to be constant within each segment. For both approaches we restricted all pairwise covariances (of the measurement-level variance-covariance matrix) to be zero.

**Challenge: model selection across multiple studies**

Applying model selection separately within each cohort can bias selection towards simple functional forms as single studies may lack power to select complex functional forms [25, 26]. Conversely, conducting model selection on the combined data from all studies may be impractical due to the large volume of data.

We considered two alternative approaches: (1) Perform model selection on the summed likelihood across the cohorts (i.e., the same model is fitted separately to each cohort and the likelihoods are summed across the cohort) [27]. (2) Perform model selection on a random sample of individuals from the combined data, where the proportion of individuals from each cohort is the same as per the combined data. An advantage of this second approach is that the
remaining data (i.e. the unselected individuals) can be used as a “validation dataset” to further evaluate model fit.

Selection criteria used to compare non-nested models include Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the mean squared prediction error (MSPE) (see Appendix for further information).

In our example, we were not able to compare natural spline modes using the summed likelihood approach due to the cohorts’ differing age ranges (e.g., not possible to fit a model with knot points positioned at the same ages to cohorts with age ranges birth to 5 years, and 9 to 11 years). Instead we selected the top two fractional polynomial and natural spline models based on a random sample of 15,000 children, and then compared the fit of these four models using the validation dataset.

To reduce the number of candidate models, model selection was conducted in two stages. At stage-1 we selected the form of the nonlinear growth trajectory, where all models included the intercept and trajectory terms as fixed effects and individual-level random effects, and the measurement-level random effects were two independent random intercepts for age-periods 0 to 2 years and >2 years. At stage-2 we selected the measurement-level covariance structure of the best-fitting model from stage-1.

**Including covariates**

In addition to the cohort variable, we included covariates child’s sex, child’s ethnicity, and parental education as fixed effects and for each covariate we included interactions between
the covariate and the trajectory terms. These fixed effects were added to the best-fitting multilevel model.

**Challenge: accounting for missing outcome data**

Likelihood estimation of multilevel models utilise all observed repeated measurements and is not biased by the missing data when differences between the observed and missing data can be explained by associations with the observed outcome and covariate data (known as missing at random [28]). In the absence of auxiliary information (i.e., data on variables not included in the multilevel model) this likelihood-based approach makes the same assumption about the missing data as standard implementations of multiple imputation (MI) and will yield standard errors with the same or greater precision [29, 28]. See the Appendix for comments on measuring missing outcome data.

For ALSPAC, the age of last measurement indicates that a sizeable proportion of the cohort exited the study prematurely, which was also reflected in the relatively wide variability in the number of measures per child (Table 1). In comparison PROBIT, which had a similar age range to ALSPAC, had lower levels of dropout and lower variability in frequency of measurement. BCG was originally a trial with a prescribed measurement schedule, which reflects the high levels of participant retention and near-constant number of measures per child. CHS was a school-based, so study participant retention was high. Despite its prescribed measurement schedule, there was sizeable variation in the frequency of measurements (lower and upper quartiles were 49 and 67, respectively). This variability was mainly due to children entering the study (i.e. enrolling at the school) at different ages with 47% of children enrolling before age 11 years, 52% between ages 11 and 12, and 1% aged 13 years or older.
Since our example did not have any auxiliary information, we decided not to impute the outcome data.

**Challenge: accounting for missing covariate data**

Combining data across cohorts raises the problem of systematically missing covariate data (i.e., covariate is missing for all participants of a cohort because it was not measured [30]). Restricting the analysis to participants with observed data on all covariates, which we shall call “complete covariate analysis” (CCA), and MI are two approaches to account for missing data. The choice of approach will depend on the missing data setting [29]. Specialised MI methods (e.g., [31, 32, 33, 34, 35]) and software (e.g. jomo [36] and Stat-JR [37]) are required when the main analysis is a multilevel model [29, 38].

In our example, there were missing data on the child-level covariates ethnicity, and both components of parental education (Table 1). The discussion below excludes the missing data on child’s ethnicity in BCG, CHS and PROBIT as we assumed all children of these cohorts were of White European origin.

We decided to use MI instead of CCA because: (1) our investigations indicated that the chance of having complete covariate data depended on the observed outcome data; thus, a CCA could be biased by omitting children with missing data [29]. (2) CCA would be extremely wasteful, discarding all data on 17.4% of children: all 2498 children of CHS and BCG, and 5734 children of ALSPAC and BiB. Factors identified as predictors of missingness, and the missing values were included in the imputation model. Since the aim of our example was to illustrate our methodology, we only considered a small number of factors.
Neither jomo nor Stat-JR were suitable for our example because they did not allow the measurement-level variance to vary with age and applying an imputation model that assumed constant measurement-level variance failed to converge. Instead we adapted an imputation procedure developed for imputing level-2 data using summary measures of any level-1 data [33]. Our adaption was an iterative procedure which incorporated summaries of the children’s growth trajectories in the imputation model. It accounted for the interactions of the main analysis because we used the same model to derive these child-specific growth trajectory summaries (see Appendix for further details). We generated 25 imputed datasets and combined the multiple sets of results into a single inference using Rubin’s rules [39].

RESULTS

The five cohorts contributed data from birth to age 20 years. The combined sample size was 47,205 children with 542,781 weight measurements (Table 1). The cohorts covered overlapping but differing age ranges with repeat weight assessments: ALSPAC 0 to 20 years, BCG 0 to 5 years, BiB 0 to 7 years, CHS 9 to 18 years, and PROBIT 0 to 18 years. CHS had the highest median number of weight measurements per child (at least quadruple that of the other cohorts), and the second highest was BCG, with a median of 14 measures.

Figure 1 shows observed mean growth trajectories among the cohorts. The general trajectory shape was nonlinear: curving over in the first year, curving slightly under between ages 5 and 12 years and starting to plateau around age 17 years. Compared to ALSPAC and PROBIT, CHS had a lower growth trajectory.

The results of the model selection procedure and a summary table of the components of the final model (Appendix-table 7) are reported in the Appendix. Figure 2 shows predicted mean
weight trajectories, using the MI point estimates of the final model, between children of different cohorts, sexes, ethnic groups, and parental education groups (in each case holding all other covariates constant). Weight trajectories were very similar between the cohorts in the first four years of life. Between ages 10-15 years children from ALSPAC were heaviest, PROBIT in between and CHS lightest (e.g., at age 15 years the predicted mean difference in weight between children from PROBIT and ALSPAC was -0.88 kg [95% confidence interval (CI) -0.44, -1.33 kg] and between children from CHS and ALSPAC was -6.83 kg [95% CI -6.17, -7.49 kg]). By age 20 years the difference in weight between children from CHS and ALSPAC had narrowed (predicted mean difference -2.84 kg [95% CI -1.57, -4.10 kg]). The marked plateauing effect after age 15 years for PROBIT could be due to its limited number of measures between 15-20 years. The weight trajectories of boys and girls were similar until adolescence and start to diverge after age 15 years. There were very little differences between children of different ethnic backgrounds or between those whose parents had different educational levels.

**DISCUSSION**

Life-course epidemiology should ideally use repeatedly assessed data from when participants were *in utero* or born through to old age. The oldest birth cohorts are from the 1950s/60s [40, 41], meaning even they only go up to mid-life. These two cohorts also lack very detailed repeat measurements even of weight and height, with substantial gaps between childhood and later adult measures. An alternative is to combine cohorts that cover different but slightly overlapping periods of the life-course to understand the impact of developmental/early life exposures on outcome trajectories across life [2]. We have illustrated the key challenges in this alternative approach and that it not only allows examination of a longer life-course but also direct comparison of the same parts of the life-course in different geographical regions.
and time periods. We have also introduced a novel approach of imputing individual-level covariates of a multilevel model with a nonlinear growth trajectory and interactions.

In order to cover a life-course it is usually necessary to combine heterogeneous cohorts, in contrast to a meta-analysis of longitudinal studies (e.g., [42]), which selects cohorts for analysis based on their similarity of design. It is also necessary to make some modelling assumptions common to all cohorts in order to gain information over separate modelling of each cohort. This requires the analyst to make decisions on which common assumptions are plausible and which notable differences between the cohorts must be accommodated in the model. In our example, the weight trajectories of each cohort had the same shape and so it was plausible to model all of the data using the same nonlinear trajectory. At the same time, we had to allow for important differences such as a lower mean trajectory for CHS compared to ALSPAC and PROBIT.

Data harmonisation is required when pooling heterogeneous data from different studies. By necessity, the level of detail in the harmonised variable is often determined by the cohort with the most simplified definition of the variable. Also, even after harmonisation the interpretation of the harmonised variable may differ between cohorts (e.g., the meaning of maternal education may change over time or geography).

In meta-analyses, MI has been proposed as a solution to systematically missing covariate data [30, 43, 44] and a study examining early and midlife risk factors for cardiovascular disease in four pooled cohorts imputed exposure variables not available in younger cohorts [45]. In our example we multiply imputed the systematically missing data on maternal education but used a simpler approach for ethnicity since population demographics at the time indicated that
almost of all of the participants would have been of White European origin. For cohorts with
greater ethnic diversity, external population-level information could be utilised to multiply
impute ethnicity as part of a sensitivity analysis [46].

Caution is needed when interpreting parameters common to all cohorts. In our example, the
trajectories of boys and girls appeared to diverge from age 15 years onwards. Whilst this may
be a real phenomenon we must also consider that this pattern may be unduly influenced by
the CHS cohort, which was the only cohort to cover the time-period 1930s to 1950s, included
data on boys only and its weight trajectories would likely differ from those of boys from later
periods due to changes in lifestyle and increases in the average height.

In summary, careful analysis can harmonise and bring together information from different
cohorts to inform life-course trajectories. Our approach can leverage several cohorts to
inform about life-course trajectories and examine heterogeneity between cohorts to shed light
on influences on trajectories.

ACKNOWLEDGEMENTS

We thank the study executives of ALSPAC, BiB, and PROBIT, Christ’s Hospital, and Dr PC
Elwood (MRC Epidemiology Unit, South Wales) and Prof. Y Ben-Shlomo (University of
Bristol) for permitting access to the data. ALSPAC, BCG study, Born in Bradford, CHS and
PROBIT were only possible because of the enthusiasm and commitment of the children and
parents. We are grateful to all the participants, health professionals and researchers who have
made these studies happen.
Funding  This study was supported by funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle) and the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement (Grant number 669545; DevelopObese). RAH, KT and DAL work in or are affiliated with a unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_00011/3 and MC_UU_00011/6). RAH is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 215408/Z/19/Z) and DAL is a National Institute of Health Research Senior Investigator (NF-0616-10102). The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and RAH, KT AND DL will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). The BCG study was funded by the UK Department of Health. BiB is supported by the Wellcome Trust (WT101597MA) a joint grant from the UK Medical Research Council (MRC) and UK Economic and Social Science Research Council (ESRC) (MR/N024397/1), the British Heart Foundation (CS/16/4/32482) and the National Institute for Health Research (NIHR) under its Collaboration for Applied Health Research and Care (CLAHRC) for Yorkshire and Humber and the Clinical Research Network (CRN). The CHS cohort was funded by Cancer Research UK and the MRC. PROBIT was originally funded by the National Health Research and Development Program (Health Canada), the Thrasher Research Fund, the United Nations Children’s Fund (UNICEF), and the European Regional Office of the World Health Organization (WHO). None of the funders influenced the content of the paper or the decision to publish. The views expressed in this paper are those of the authors and not necessarily any funder listed here.
REFERENCES


[34] R. Wijesuriya, M. Moreno-Betancur, J. B. Carlin and K. J. Lee, “Evaluation of


Box 1. Key challenges

Data harmonisation
- Information may be lost during the data harmonisation process in order to include cohorts with a relatively crude measure of a variable.
- Some variables may not be measured by all cohorts. Subject-matter knowledge may provide information about their most likely values.

Accounting for the data’s dependence structure
- The data’s dependence structure must be appropriately modelled in order to obtain appropriate standard errors.
- Reliable estimation of the random effects’ variance components at a particular level depends on the number of units at that level and not the total number of level-1 observations (e.g., repeated measurements).

Model selection across multiple cohorts
- Model selection on the combined data from all studies may be impractical due to the large volume of data.
- Two alternatives are: (1) fit the same model separately to each cohort and perform model selection on the summed likelihoods across the cohorts. (2) Perform model selection on a random sample of individuals, keeping the proportion of individuals from each cohort as per the combined data.
- Model selection based on summed likelihoods may not be feasible when the feature under comparison (e.g., nonlinear trajectory) requires each cohort to cover the same age range (e.g., restricted cubic splines).

Missing data in the outcome (repeated measurements) and the covariates
- Determining the amount of missing outcome data is difficult for cohorts without a prescribed measurement schedule (such as opportunistic health visits).
- Likelihood estimation of a multilevel model can utilise all available observed outcome data and is valid under the missing at random assumption (i.e., differences between the observed and missing data are explained by associations with the observed outcome and covariate data).
- When outcome data are suspected missing not at random, multiple imputation (MI) can use information from auxiliary variables (not included in the multilevel model) to explain the reasons for the missing data and provide valid inference.
- Restricting the analysis to those with complete covariate data can result in a substantial loss of information especially when covariates are systematically missing in some cohorts (missing for all individuals) and/or when there are many covariates each with small amounts of missing data.
- The imputation model of MI should account for the main features of the multilevel model (e.g., its multilevel structure, any interactions).
Table 1: Baseline characteristics of the cohorts using harmonised categories

<table>
<thead>
<tr>
<th>Region</th>
<th>ALSPAC</th>
<th>BCG</th>
<th>BiB</th>
<th>CHS</th>
<th>PROBIT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>South West England</td>
<td>South East Wales</td>
<td>Centre North of England</td>
<td>South East England</td>
<td>Republic of Belarus</td>
</tr>
<tr>
<td>Age range</td>
<td>0-20</td>
<td>0-5</td>
<td>0-6</td>
<td>9-18</td>
<td>0-16</td>
</tr>
<tr>
<td>No. participants(^a)</td>
<td>14,216</td>
<td>951</td>
<td>13,445</td>
<td>1,547</td>
<td>17,046</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>51%</td>
<td>54%</td>
<td>52%</td>
<td>100%</td>
<td>52%</td>
</tr>
<tr>
<td>female</td>
<td>49%</td>
<td>46%</td>
<td>48%</td>
<td>0%</td>
<td>48%</td>
</tr>
<tr>
<td>missing</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>79%</td>
<td>100%(^b)</td>
<td>39%</td>
<td>100%(^b)</td>
<td>100%(^b)</td>
</tr>
<tr>
<td>South Asian</td>
<td>0%</td>
<td>0%</td>
<td>50%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Other</td>
<td>4%</td>
<td>0%</td>
<td>8%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>missing</td>
<td>17%</td>
<td>0%(^c)</td>
<td>3%</td>
<td>0%(^c)</td>
<td>0%(^c)</td>
</tr>
<tr>
<td>Maternal education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left school at 15 or 16</td>
<td>55%</td>
<td>0%</td>
<td>43%</td>
<td>0%</td>
<td>4%</td>
</tr>
<tr>
<td>left school at 17 or 18</td>
<td>19%</td>
<td>0%</td>
<td>12%</td>
<td>0%</td>
<td>82%</td>
</tr>
<tr>
<td>degree</td>
<td>11%</td>
<td>0%</td>
<td>21%</td>
<td>0%</td>
<td>14%</td>
</tr>
<tr>
<td>missing</td>
<td>15%</td>
<td>100%</td>
<td>24%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Paternal occupation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>class I or II</td>
<td>22%</td>
<td>17%</td>
<td>13%</td>
<td>55%</td>
<td>11%</td>
</tr>
<tr>
<td>class III</td>
<td>36%</td>
<td>58%</td>
<td>19%</td>
<td>21%</td>
<td>60%</td>
</tr>
<tr>
<td>class IV, V or other</td>
<td>21%</td>
<td>23%</td>
<td>44%</td>
<td>3%</td>
<td>25%</td>
</tr>
<tr>
<td>missing</td>
<td>21%</td>
<td>2%</td>
<td>24%</td>
<td>21%</td>
<td>4%</td>
</tr>
<tr>
<td>Total no. weight measures</td>
<td>157,000</td>
<td>12,737</td>
<td>78,110</td>
<td>89,070</td>
<td>205,864</td>
</tr>
<tr>
<td>Median no. measures per child (IQR)</td>
<td>10 (10)</td>
<td>14 (1)</td>
<td>5 (3)</td>
<td>57 (18)</td>
<td>13 (5)</td>
</tr>
<tr>
<td>Median ALM(^c) in years (IQR)</td>
<td>13.8 (13.1)</td>
<td>5 (0)</td>
<td>4.7 (2.7)</td>
<td>17.8 (1.5)</td>
<td>16.0 (1)</td>
</tr>
</tbody>
</table>

\(^a\): Included in the growth analysis. \(^b\): These cohorts had no record of ethnicity. Based on the populations from which they were recruited we assigned all participants to White European ethnicity. \(^c\): Age at last measurement.
**Figure 1:** Observed mean weight trajectories among cohorts
Figure 2: Predicted mean weight trajectories according to (a) cohort study, (b) child’s sex, (c) child’s ethnicity, and (d) parental education. The solid black line in each graph denotes the predicted mean weight trajectory for the reference participant: white boys from the ALSPAC cohort whose mother’s left school aged 15 or 16 from the ALSPAC cohort and father’s occupation is classified as professional or managerial.