

1 **ORIGINAL ARTICLE**

2 TITLE: Erosion of representativeness in a cohort study

3

4 AUTHORS: Christodoulou, M.D.<sup>1\*</sup>, Brettschneider, J.A.<sup>2</sup>, and Steinsaltz, D.<sup>1</sup>

5

6 <sup>1</sup> Department of Statistics, University of Oxford, Oxford, U.K.

7

8 <sup>2</sup> Statistics Department, University of Warwick, Coventry, U.K.

9

10 \* For correspondence e-mail: [maria.christodoulou@stats.ox.ac.uk](mailto:maria.christodoulou@stats.ox.ac.uk)

11

12

13 **SUMMARY:**

14

15 Background: The National Child and Development Study (1958 British Birth Cohort) follows the  
16 lives of over 17 000 people born in a single week in England, Scotland, and Wales. Since the  
17 initial recruitment there have been nine sweeps to gather subsequent life-course data, and a  
18 Biomedical Sweep (Age 44) – between Sweeps 6 (Age 42) and 7 (Age 46) – that has found  
19 widespread application in genetic studies. Due to its non-selective recruitment, the survey is  
20 frequently used as a representative proxy for the British population in demographic,  
21 epidemiological, and medical studies. We examine the effect of attrition on representativeness  
22 of female fertility and education length.

23

24 Methods: We compare numbers and timings of fertility-related events of female cohort  
25 members with national estimates. Spline approximation was used to link records with different  
26 aggregation intervals. Participants present in the Biomedical Sweep (Age 44) were compared  
27 to those who were not.

28

29 Results: We established that both timings and counts of maternities and terminations in the  
30 cohort diverge from the patterns of their contemporaries. For women who participated in the  
31 Biomedical Sweep (Age 44), we noted positive correlations of study continuation with years  
32 spent in full time education, and with age at first birth. We determined that women who did  
33 not participate in the Biomedical Sweep (Age 44) reported different fertility patterns from those  
34 who did.

35

36 Conclusions: It is possible to use National Statistics to quantify various forms of selection bias  
37 that inevitably creep into even the most meticulously sampled longitudinal study, and the  
38 misreporting that affects particular questions. While the 1958 Birth Cohort has reasonably been

39 described as “broadly representative” of the British population, such a characterisation  
40 becomes questionable when the data are to be applied to questions about female fertility.

41

42 Keywords: Bias, attrition, representativeness, fertility, age at menarche, age left full time  
43 education, birth cohort, 1958 British Birth Cohort, NCDS, dropouts.

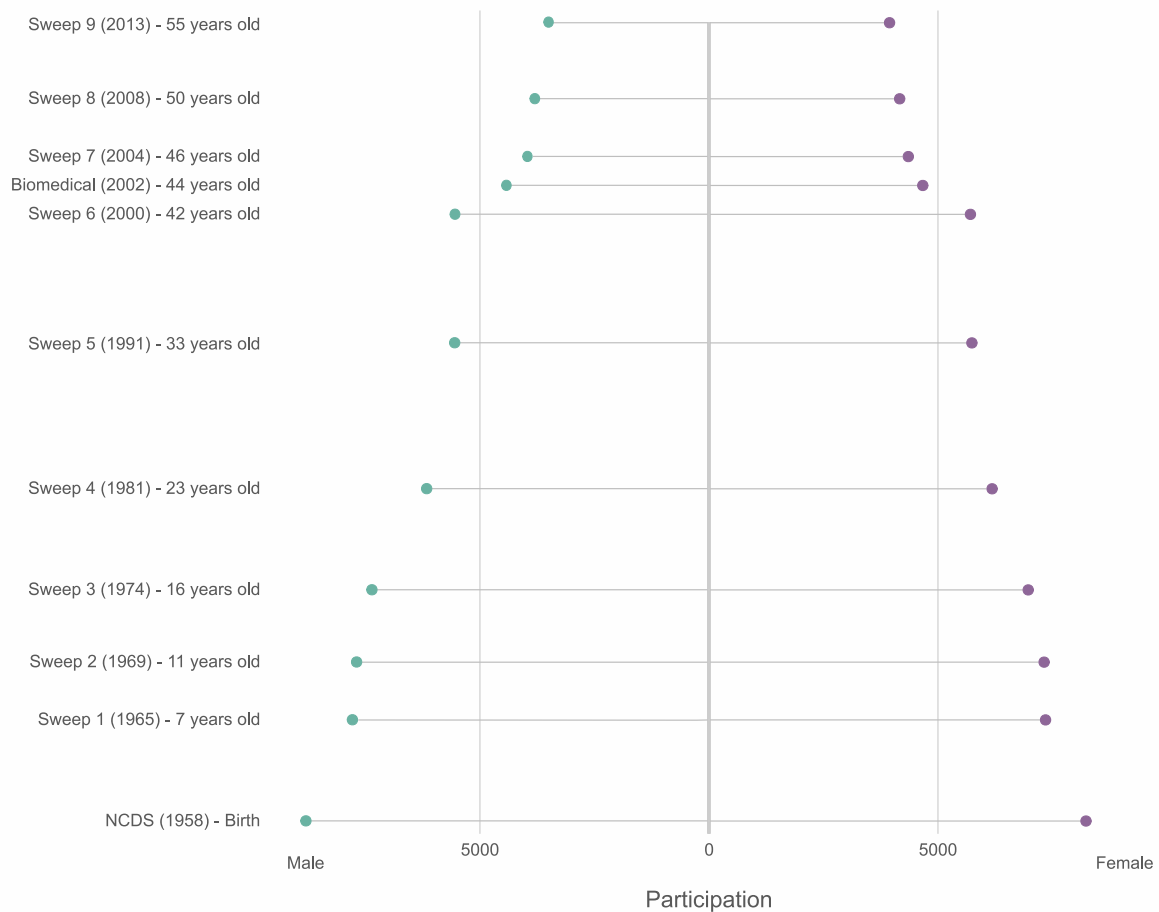
44 **KEY MESSAGES:**

- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60
- Fertility patterns of women reported in later sweeps of the 1958 British Birth Cohort diverge from those of their contemporaries as estimated from national statistics using spline approximation. Both maternities and terminations are underreported.
  - Female participants who dropped out earlier follow national maternity trends more closely than those who participated in the study longer, but they still report slightly fewer maternities than national statistics suggest.
  - Female participants who persisted through later sweeps experienced first births later and left education at a later age than those who dropped out earlier.
  - Although the 1958 British Birth Cohort has been judged as representative of the British population for some research questions, the cohort population presents increasingly biased fertility patterns in female participants over time. Studies related to fertility using data from this cohort may require adjustment.

## 61 **BACKGROUND**

62 The 1958 British Birth Cohort (National Child Development Study) was originally designed as a  
63 one-off study focusing on factors surrounding perinatal mortality in Great Britain, and including  
64 all births occurring in England, Scotland, and Wales during the first week of March 1958 (1).  
65 Cohort members have been followed throughout their lives, creating a rich dataset with  
66 extensive records on multiple aspects of their lives, such as health, education, and  
67 employment. The non-selective inclusion of every child born in Great Britain in one week has  
68 made it an attractive starting point for longitudinal studies of a representative cross-section of  
69 the British population. A large and diverse collection of variables has been amassed for the 17  
70 415 initial cohort members (and the 925 added in the replenishment sweeps at ages 7, 11, and  
71 16 for children immigrating to Great Britain), documenting every facet of their lives, from  
72 gestation through late middle age. Despite dropouts (due to mortality, emigration, or other  
73 reasons), more than half of the cohort (9 377 members) persisted to the Biomedical Sweep  
74 (age 44), and more than 40% participated in the most recent sweep in 2013. Levels of attrition  
75 are illustrated in Figure 1.

76



77

78 Figure 1: Male and female productive participation by sweep for the 1958 British Birth Cohort  
 79 (National Child Development Study – NCDS). Starting with 17 415 cohort members at  
 80 recruitment (8 411 female productive participants and 9 004 male), productive participation  
 81 declined to 9 377 members for the Biomedical Sweep (4 712 female participants and 4 665  
 82 male) in 2002. Age of cohort member and year of Sweep are listed as the start year for each  
 83 sweep. The highest average annual rate of loss occurred between Sweep 6 and the  
 84 Biomedical Sweep. This pattern is consistent for both female and male participants. Productive  
 85 participants are the cohort members who were successfully contacted and responded to a  
 86 Sweep's questions. As only female participants were released for this study, the male  
 87 participation was calculated using the publicly available achieved samples for the cohort. The  
 88 lengths of the lines are proportional to the numbers of individuals they represent.

89

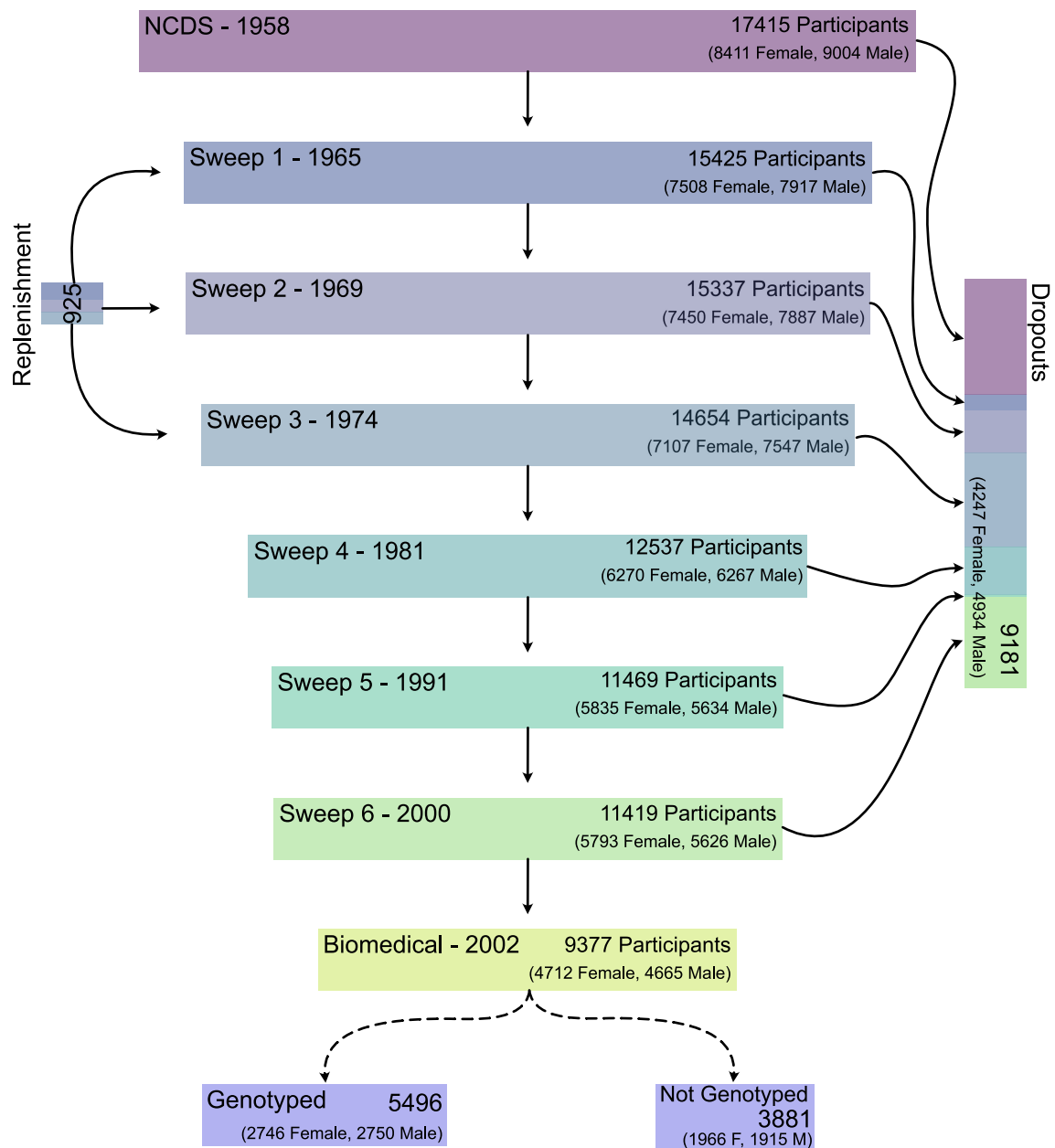
90 Because of the outsized role that this cohort has played in a wide range of research, the  
91 development of its membership over time has received considerable attention. Atherton et al.  
92 (2) compared a variety of sociodemographic, health, and behavioural variables between  
93 Sweep 1 (Age 7) and the Biomedical Sweep (Age 44). They concluded that subject attrition  
94 had not substantially vitiated the broadly representative composition of the persisting cohort.  
95 This claim has been echoed in numerous publications referring to the cohort as “nationally  
96 representative” (3–9) or “broadly representative”, subject to attrition (10–16).

97

98 The present work examines whether specific data biases may distort the picture of particular  
99 research questions within an otherwise representative sample. Our focus is on the impact of  
100 attrition on fundamental indicators of education and fertility, such as age at menarche and  
101 years spent in education. We explore associations between these variables and the length of  
102 continuing participation in the study, for women present in the Biomedical Sweep (Age 44).  
103 We further compare national statistical records to the maternity and termination reports of  
104 female participants, separately for those who persisted until the Biomedical Sweep (Age 44)  
105 and those who did not. Our objective for this is to establish how the two groups compare to  
106 one another, and whether they diverge in parallel from national statistics, or follow predictable  
107 patterns. As the Biomedical Sweep (Age 44) was focused on collecting biological material and  
108 clinical data, it is only participants in this Sweep whose information is routinely released for  
109 research that aims to include genetic data. The genetic data commonly used were initially  
110 generated by the Wellcome Trust Case Control Consortium (WTCCC) studies (17,18) (Figure 2),  
111 which used members of the 1958 Birth Cohort as part of the control group for large meta-  
112 analyses on common diseases. These were subsequently made available to other researchers,  
113 and have since been used as part of control groups (usually in combination with UK Blood  
114 Donors) for a variety of medical studies, including work on leukaemia (19,20), schizophrenia  
115 (21) and rheumatoid arthritis (22), as well as for sociological and demographic research (23,24).  
116 As these studies include only participants in the Biomedical Sweep (Age 44), we take as our  
117 primary cutoff. Our crucial aim is to better understand the composition of this group, that has

118 served as a control in so many medical studies, and how it compares to the remainder of the  
119 cohort.  
120





121

122 Figure 2: Male and female productive participation, dropouts, and replenishment in the 1958

123 British Birth Cohort until the Biomedical Sweep, culminating in the numbers of individuals that

124 were genotyped. Productive participants are the cohort members who were successfully

125 contacted and responded to a sweep's questions. As only female participants were released  
126 for this study, the male participation was calculated using the publicly available achieved  
127 samples for the cohort. Lengths of rectangles are proportional to the total number of individuals  
128 participating.

129

130 Whether representativeness as a concept matters for epidemiological, sociological, or  
131 demographic research has been extensively debated (25–30), and is beyond the scope of this  
132 work. We are not questioning whether overall the 1958 British Birth Cohort is a representative  
133 proxy of British population for its age group. Nor are we questioning the appropriateness of the  
134 cohort members specifically as a control group for the genetic studies designed by WTCCC.  
135 Here we consider the essential yet sometimes neglected distinction between a population  
136 being representative overall, versus being representative for a particular research objective.

137 In this work we achieve two aims. First, we examine the representativeness of the cohort in  
138 matters of female fertility, which is interesting on its own merit as well as a fundamental vital  
139 statistic that can be verified against National Registry Data. We use education as a  
140 sociodemographic variable, as a plethora of studies have demonstrated the links between  
141 education and fertility behaviour(31–34). Second, we model how researchers can benchmark  
142 a study against available national statistics in order to quantify the divergence between the  
143 studied group and their contemporaries.

144

145 We have instrumentalised the general question of representativeness in two specific questions:

- 146 1. whether annual rates of maternity and termination reported by participants who  
147 persisted through the later sweeps diverged from rates reported by those who dropped  
148 out earlier, and how each of these comports with national statistics. Due to different  
149 age aggregation in the latter, we used spline approximation to estimate annual rates,  
150 described in the Section: "Annual maternity and termination rate estimates"; and

151           2. whether participant attrition interacts with education and fertility variables. This was  
152           conducted using time-to-event analysis, and is presented in the Section: "Attrition  
153           through the sweeps".

154   Combined, these delineate the impact of attrition on the overall makeup of the cohort for  
155   fundamental fertility and education variables. It should be noted that miscarriages have not  
156   been examined, as no comparable national statistics are available.

157

158

## 159 **METHODS**

### 160 **DATA AND SOFTWARE**

161 All available fertility-related variables from the 1958 Birth Cohort were requested from  
162 METADAC (Managing Ethico-social, Technical and Administrative issues in Data Access) and  
163 used to reconstruct in-depth fertility profiles for the 4 712 female participants in the Biomedical  
164 Sweep (Age 44) conducted in 2002. We denote by “year of Sweep” the year in which a given  
165 sweep started. Protocols for clean-up (the accuracy of which was tested using synthetic data)  
166 are supplied in the **Supplementary Material**. Variables relating to participation in sweeps, and  
167 age of leaving full-time education were also requested. A list of all 942 requested variables is  
168 provided in the **Supplementary Material**. As we were interested in the impact of attrition, we  
169 further requested information about the female cohort members who did not participate in  
170 the Biomedical Sweep (Age 44). This yielded 4 247 additional participants, with 314 variables  
171 requested (a subset of the original variable request, relating to the available sweeps). No data,  
172 if available were released for Sweep 6 (Age 42) for this group. Although we noted some slight  
173 discrepancies between the numbers of participants in our data and those recorded on  
174 published tables in Atherton et al.(2) and Power et al (1), we are in overall agreement on the  
175 productive participant counts. As part of our quality-control assessment, we noted that initial  
176 recruitment sex ratio (107.05 males to 100 females) matches the birth sex ratio of the general  
177 population. Further consistency checks are detailed in the associated R scripts.

178

179 Maternities (live births and stillbirths) and terminations as defined by the Office for National  
180 Statistics for all women living in Great Britain born in 1958 were obtained from the Office for  
181 National Statistics (ONS) and from National Records of Scotland (NRS) (35), for all years  
182 between 1972 and 2001. The ONS definitions can be found in full in ONS maternity and  
183 termination records (covering England and Wales) prior to 1992 were available only as age-  
184 group aggregates through the Conception Statistics tables (36,37). From NRS both age-group  
185 aggregates and a breakdown by mother's age were available (35). The life table for women

186 born in 1958 was obtained from the Human Mortality Database (38). Statistical analysis was  
187 conducted using base R (39), “tidyverse” (40), and “survival” (41) through RStudio (42). All code  
188 is available in the **Supplementary Material**.

189

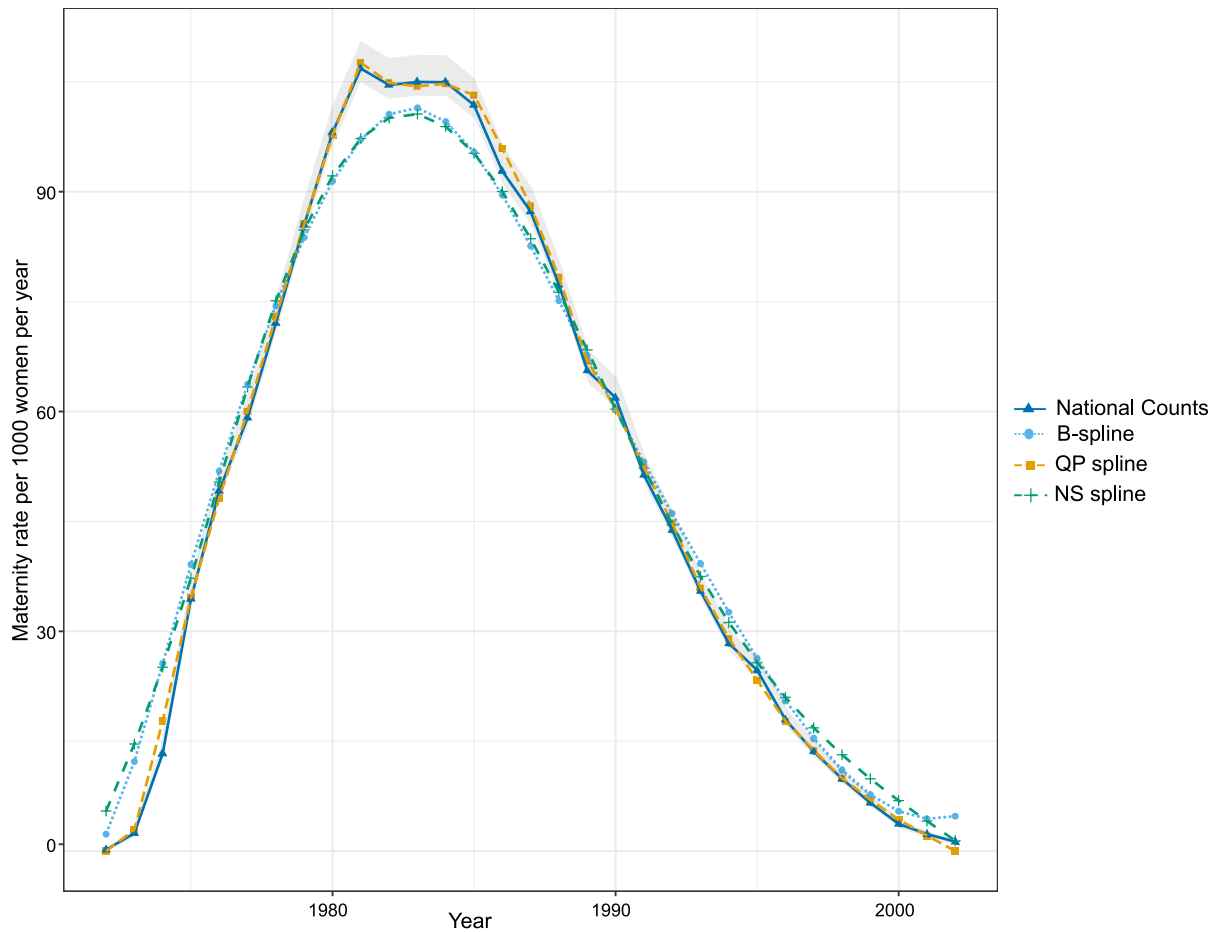
## 190 **ANNUAL MATERNITY AND TERMINATION RATE ESTIMATES**

191 We compared rates of maternities and terminations reported by participants with the  
192 corresponding rates in the general population, derived from official statistics. Although counts  
193 by individual years of mother’s age are available from the Scottish data, we aggregated the  
194 age groups to align them with the data releases from England and Wales. To then extract  
195 single-year estimates of maternity and termination rates for individuals born in 1958 we fitted  
196 splines, for which we tested three different specific procedures: a standard b-spline (knots at  
197 1974, 1978, 1983, 1988, and 1993 – corresponding to the years in which participants crossed  
198 over between age classes of the vital statistics tables, for example 1974 when they entered the  
199 16-19 years age class – a natural cubic spline (knots as in the b-spline), and a quadratic  
200 optimisation spline proposed by Grigoriev et al. (43). This last is based on minimising the sum of  
201 squared second-order differences, yielding annual rate estimates (per 1000 women) for  
202 maternity and terminations. These were compared with the per-1000-women rates for  
203 maternities and terminations within the 1958 Birth Cohort, for which we also calculated 95%  
204 confidence intervals.

205

## 206 **SPLINE VERIFICATION**

207 We tested the ability of the spline methodology to recover the known individual-year counts  
208 for Scottish maternities from multi-year aggregated Scottish maternity data. We transformed  
209 the Scottish birth counts into rates per 1000 women by using the life table from the Human  
210 Mortality Database. These are plotted together with Scotland-only spline estimates in Figure 3.  
211 As the quadratic-programming spline fitted the count rates better than the natural or b-splines,  
212 we rely primarily on this method in our subsequent analysis.



214

215 Figure 3: Maternity spline estimates extracted from age group aggregates against single-year  
 216 counts for Scotland. Three spline methods are presented: B-splines (light blue), Quadratic  
 217 programming splines (yellow), and Natural cubic splines (green). These are compared with  
 218 rates per 1000 women estimated from counts (dark blue, with 95% confidence interval in light  
 219 grey).

220 **ATTRITION THROUGH THE SWEEPS**

221 We distinguished between two broad categories of participation: participating in the  
 222 Biomedical Sweep (Age 44) ("Persisters") or not ("Dropouts"). We define two different measures  
 223 of participation: the last sweep of having participated in any way ("Overall Continuation  
 224 Length") and the last sweep of having responded to fertility-related questions ("Fertility  
 225 Continuation Length").

226 Table 1: Glossary of terminology used, defining participation indicators (Persisters vs Dropouts)  
 227 and measures of Continuation Length (Overall vs Fertility).

PARTICIPATION INDICATORS	<b>Persister:</b> cohort member who participated in the Biomedical Sweep.
	<b>Dropout:</b> cohort member who did not participate in the Biomedical Sweep.
MEASURES OF CONTINUATION LENGTH	<b>Overall Continuation Length:</b> last sweep in which cohort member has participated in any way
	<b>Fertility Continuation Length:</b> last sweep in which cohort member has responded to any fertility-related questions

228

229 Although we cannot quantify possible refusal to respond to fertility questions when examining  
 230 fertility-related questions, we can at least use this distinction in participation to detect patterns.

231 **WITHIN THE PERSISTERS**

232 We examined associations between fertility and education event times: *age at menarche*,  
 233 *age at first birth*, and *age left full time education*. Age at menopause was not studied as most  
 234 of the sweeps occurred prior to the menopause age range. Kendall's tau was used as a  
 235 correlation measure between ages at dropout and at event. Analyses were conducted using  
 236 both Continuation Length measures.

237

238 For individuals lacking event times, censoring time was determined based on auxiliary  
 239 information present in the dataset. For example, for *age at first birth*, individuals with no  
 240 recorded maternities were considered censored at the last time when a birth could have been  
 241 observed using other fertility variables where available (such as menopause age). Where  
 242 insufficient information was available to make a clear judgement about censoring time, these  
 243 individuals were excluded entirely from the analysis. The numbers of excluded Persisters are  
 244 summarised in Table 2.

245 Table 2: Numbers of Persisters that were included and excluded in the analyses for *age at*  
 246 *menarche, age at first birth and age left full time education.*

	<i>Age at menarche</i>		<i>Age at first birth</i>		<i>Age left full time education</i>	
	Excluded	Included	Excluded	Included	Excluded	Included
Overall Continuation Length	0	4712	414	4298	239	4473
Fertility Continuation Length	36	4676	438	4274	275	4437

247

248 **PERSISTERS VS DROPOUTS**

249 We compared *age at menarche* and *age at first birth* between the 4 712 Persisters and the  
 250 4 247 Dropouts using log-rank tests. *Age left full time education* was not studied here as the  
 251 relevant variables were only collected in Sweep 6 (Age 42), when many Dropouts had already  
 252 left the study. Age at menopause was not studied because the last records for Dropouts would  
 253 be prior to age 44, which is before most women would have experienced menopause.

254

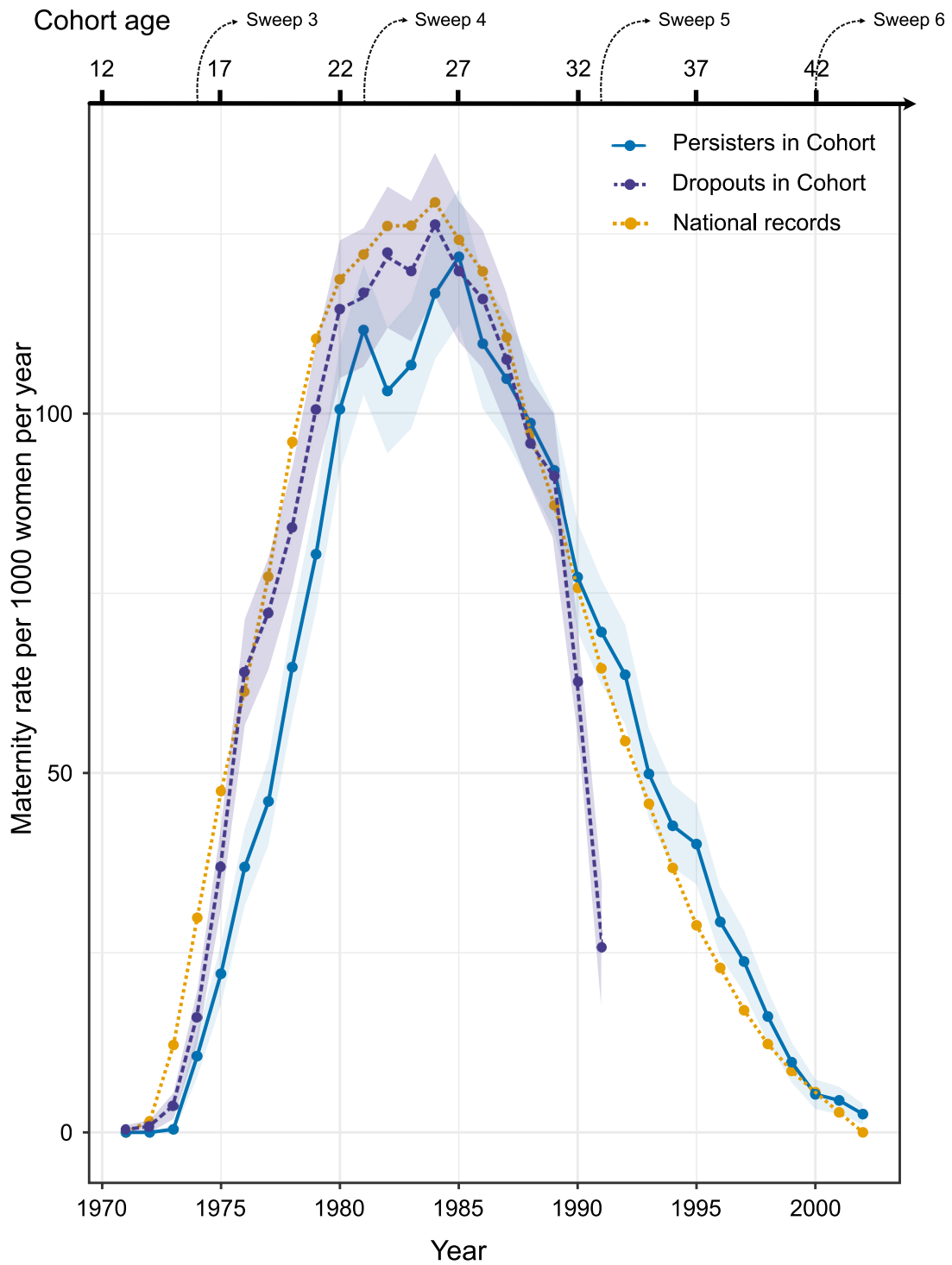


255 **RESULTS**

256 **ANNUAL MATERNITY AND TERMINATION RATE ESTIMATES**

257 Annual maternity rates for Persisters and Dropouts in the Birth Cohort, as well as estimates of  
258 these rates from national statistics, are shown in Figure 4. Until 1987, both Persisters and Dropouts  
259 show lower rates than national statistics, but Dropouts follow national statistics more closely  
260 than Persisters. Fewer maternities were recorded for Persisters between 1973 and 1984 than  
261 would be predicted from the records of the general population, with an average of 2.21%  
262 fewer maternities per year (sd: 0.73%, range: 1.01% - 3.13%). In 1982 and 1983 the deviation is  
263 particularly large, and shows a prominent dip in Persisters' maternity rate curve, separating  
264 them even more from Dropouts. From 1988 onwards Persisters' maternity rates exceed the rates  
265 calculated from national statistics.

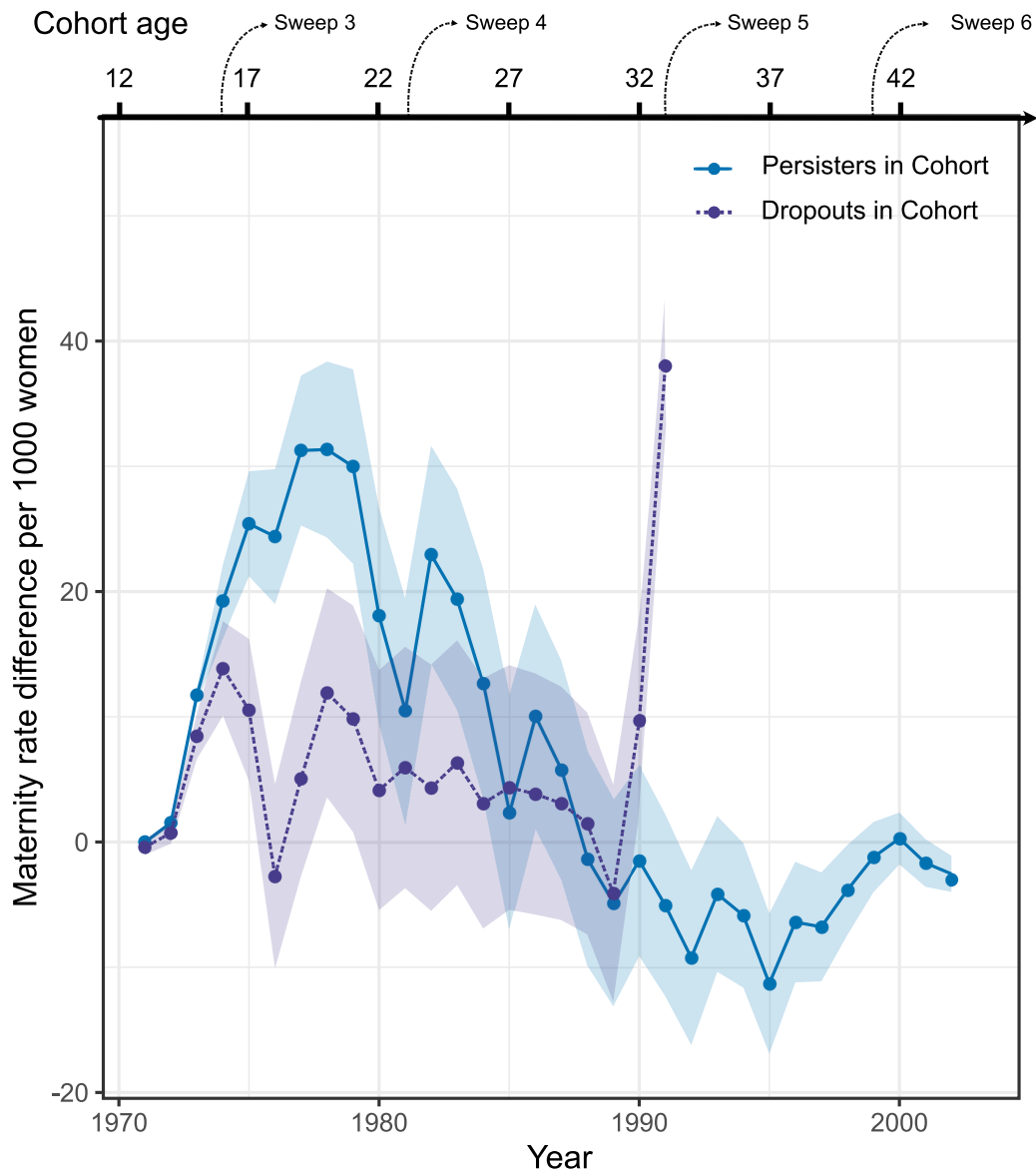
266



267

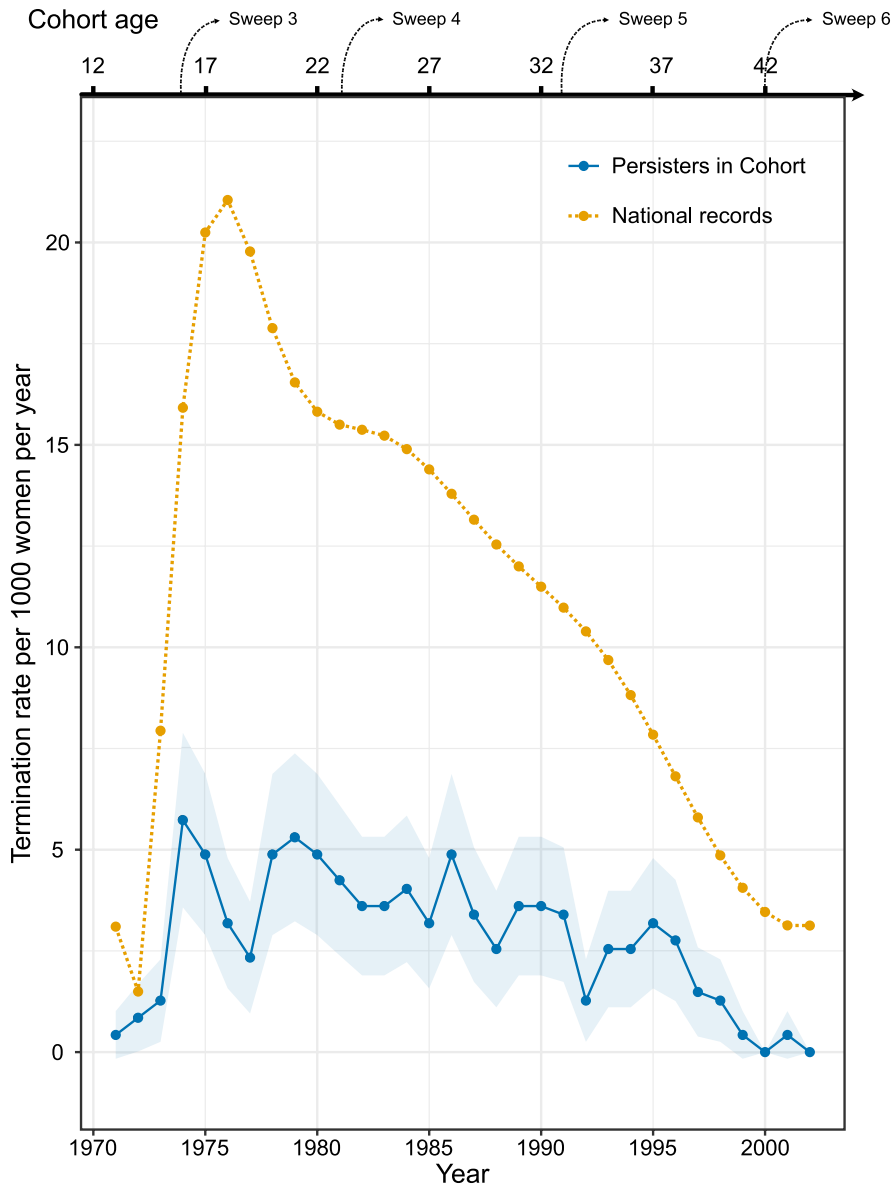
268 Figure 4: Annual maternity rates (per 1000 women) for Persisters (dark blue, with 95% confidence  
 269 interval in lighter shade) and Dropouts (purple, with 95% confidence interval in lighter shade)  
 270 in the Birth Cohort. These are compared to quadratic programming spline approximations of  
 271 these rates based on national statistics age group aggregates (yellow).

272 Maternity rate differences between the cohort participants and National statistics are  
 273 presented in Figure 5.  
 274



275  
 276 Figure 5: Annual maternity rate difference (per 1000 women) between National statistics and  
 277 Persisters (dark blue, with 95% confidence interval in lighter shade), and National statistics and  
 278 Dropouts (purple, with 95% confidence interval in lighter shade). For National statistics, rates  
 279 were estimated using quadratic programming spline approximations of age group  
 280 aggregates.

281 Comparison of termination rates between national statistics and those reported by Persisters is  
282 illustrated in Figure 6. The reported termination rates for Persisters diverge substantially from  
283 those approximated from the national statistics data. The terminations for Dropouts have not  
284 been included, as the events were not linked with specific years. This is due to the formulation  
285 of termination-related questions prior to 1990s.  
286



288

289 Figure 6: Annual termination rates (per 1000 women) as recorded for Persisters in the Birth  
 290 Cohort (dark blue, with 95% confidence interval in lighter shade) compared to quadratic  
 291 programming spline approximations of these rates based on national statistics age group  
 292 aggregates (yellow).

293 **ATTRITION THROUGH THE SWEEPS**

294 **WITHIN PERSISTERS**

295 Kendall's tau between Continuation Length past the Biomedical Sweep (Age 44) and three  
296 event times – *age at menarche*, *age at first birth*, and *age left full time education* – are  
297 summarised in Table 3. If measured by "Overall Continuation Length", the length of  
298 continuation evinces correlation with each of them – weakly negative for *age at menarche*,  
299 positive for *age at first birth* and *age left full time education*. In other words, members who  
300 participated in later sweeps experienced menarche at younger ages, and they both  
301 experienced their first birth and left full time education at later ages. If measured by "Fertility  
302 Continuation Length", the negative correlation of Continuation Length with *age at menarche*  
303 disappears, but the positive correlations with the other two event times persists. This suggests  
304 that findings from the Biomedical Sweep (Age 44) on basic fertility and education characters  
305 may not be applicable to the whole cohort.

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320 Table 3: Kendall's tau rank correlation between Continuation Length and three event times:  
 321 *age at menarche, age at first birth, and age left full time education*.  $\tau$  estimates, their 95%  
 322 confidence intervals, and the associated p-values were calculated for two participation  
 323 indicators (Overall Continuation Length and Fertility Continuation Length) resulting in similar  
 324 results for *age at first birth* and *age left full time education*, but different results for *age at*  
 325 *menarche*.

	<i>Age at menarche</i>	<i>Age at first birth</i>	<i>Age left full time education</i>
Overall Continuation Length	$\tau = -0.033$ 95% CI: (-0.045, -0.021) p-value=0.009	$\tau = 0.124$ 95% CI: (0.112, 0.136) p-value < $2.2 \times 10^{-16}$	$\tau = 0.117$ 95% CI: (0.106, 0.127) p-value < $2.2 \times 10^{-16}$
Fertility Continuation Length	$\tau = -0.003$ 95% CI: (-0.02, 0.013) p-value=0.778	$\tau = 0.137$ 95% CI: (0.12, 0.154) p-value < $2.2 \times 10^{-16}$	$\tau = 0.138$ 95% CI: (0.122, 0.153) p-value < $2.2 \times 10^{-16}$

326

327

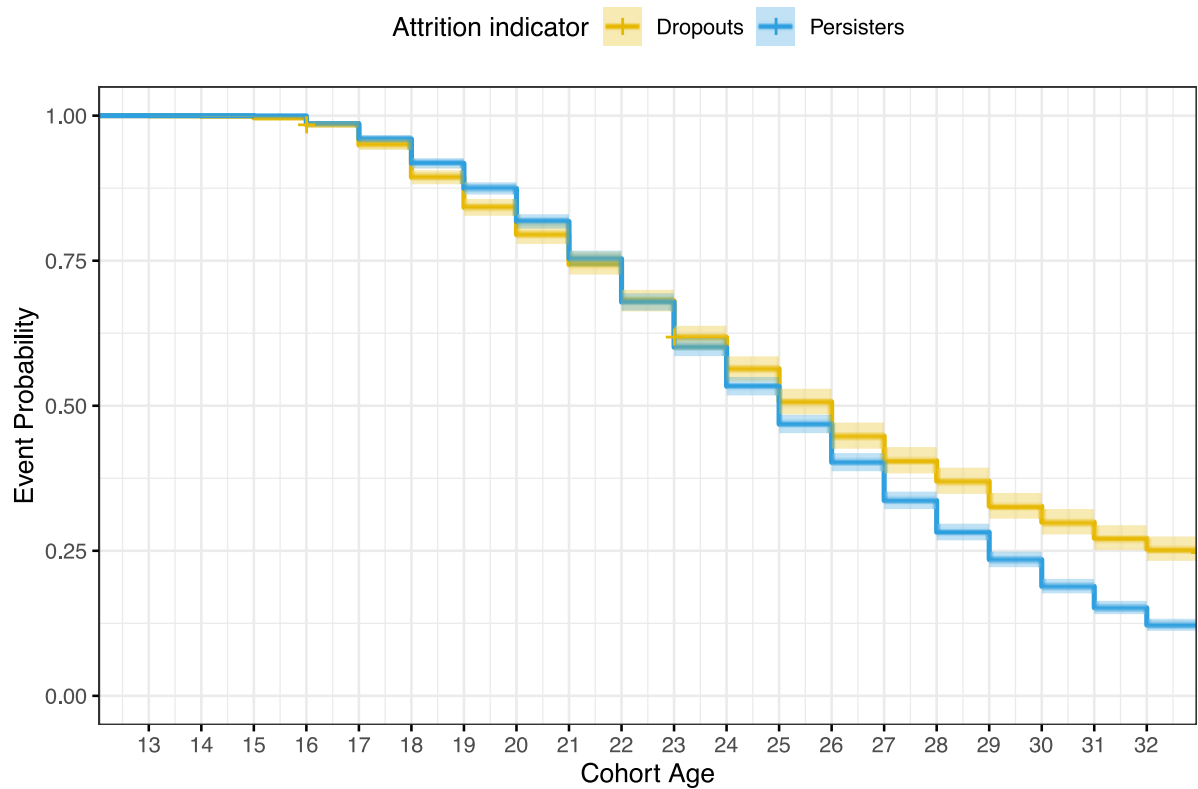
328 **PERSISTERS VS DROPOUTS**

329

330 Results of the comparison for *age at first birth* between Persisters and Dropouts in the Birth  
 331 Cohort are presented in Figure 7. Dropouts experienced an earlier age at first birth, with median  
 332 age 22, as opposed to median age 25 for Persisters.

333

## Age at First Birth



334

335 Figure 7: Kaplan-Meier curves for age at first birth comparison between Dropouts in the Birth  
336 Cohort (yellow) and Persisters in the Birth Cohort (in blue).

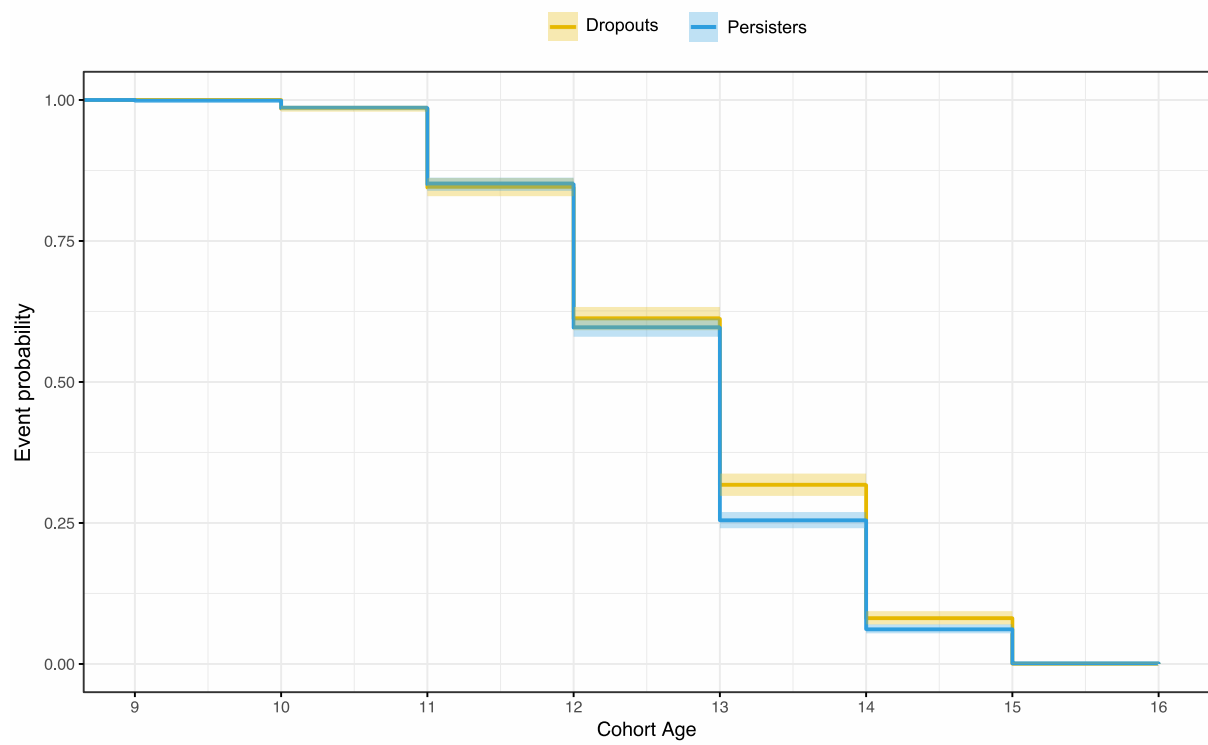
337

338 Comparison of *age at menarche* between Persisters and Dropouts in the Birth Cohort is  
339 presented in Figure 8. The two curves are similar, although proportionally more participants in  
340 the group of Persisters experience menarche at age 13.

341



### Age at Menarche



342

343 Figure 8: Kaplan-Meier curves for age at menarche comparison between Dropouts in the Birth

344 Cohort (yellow) and Persisters in the Birth Cohort (blue).

345

## 346 DISCUSSION

347 Our distinction between Persisters (women present in the Biomedical Sweep (Age 44)) and  
348 Dropouts (women who did not participate in the Biomedical Sweep (Age 44)) seems to be a  
349 productive division for understanding differential continuation rates. Suitable population  
350 estimates are often unobtainable, making direct comparisons between a study cohort and  
351 the target population impossible (44). We confirmed the utility of the quadratic-optimisation  
352 splines for recovering annual rates from events tabulated by multiyear age aggregates.  
353 We have established two key findings about the impact of attrition on basic fertility and  
354 education characteristics for the women of the 1958 Birth Cohort.

- 355 1. Fewer fertility events are reported in the cohort than would be expected from national  
356 statistics. For maternities this discrepancy lingers into the mid 1980s for Persisters and, to  
357 a lesser degree, also for Dropouts. Reported terminations for Persisters are extremely low  
358 throughout, as compared with the known rates from national statistics.
- 359 2. For Persisters both *age at first birth* and *age left full time education* are positively  
360 correlated with subjects' length of continued study participation. Persisters  
361 experienced their first birth at a later age than Dropouts. *Age at menarche* was similar  
362 overall between the two groups.

363

364 These findings suggest that Persisters have different fertility and educational profiles to  
365 Dropouts. Furthermore, this profile diverges from national trends of their contemporaries for  
366 both maternities and terminations. Because of the nature of its initial recruitment, the 1958 Birth  
367 Cohort is often considered a representative sample of the British population. Our findings  
368 suggest that this perception is not entirely accurate, at least with regard to the fertility events  
369 considered here. Following the *Strengthening the Reporting of Observational Studies in*  
370 *Epidemiology (STROBE) statement* (45) we recommend that this disparity be mentioned in  
371 future work relying on the representativeness of the cohort for fertility questions.

372

373 By comparing the maternity patterns of Persisters with National Statistics we established that  
374 there were fewer maternities reported by the cohort participants between 1973 and 1984  
375 (Figure 4). The disparity between the 1958 Birth Cohort fertility patterns and the estimates based  
376 on national statistics is perplexing. We had expected self-reported births to yield numbers and  
377 age patterns very close to those registered by statistics agencies. The discrepancy suggests  
378 either a reliability issue with the maternity records for Persisters in the Birth Cohort –  
379 underreporting or recording errors – or a genuine divergence in behaviour between Persisters  
380 and the general population. While less pronounced, Dropouts also exhibit lower maternities  
381 rates than national statistics suggest. Explanations for the observed differences need to address  
382 both the smaller numbers overall and the additional gaps due to non-random attrition. Of some  
383 interest is also the slight decrease in maternity rates for Persisters in 1982 and 1984. These are  
384 the years following Sweep 4 (Age 23), making it possible for participants to have mistakenly  
385 assumed during Sweep 5 (Age 33) that they had already reported a particular birth. This  
386 explanation does not differentiate between Dropouts and Persisters, however. As this particular  
387 dip is not observed in the Dropouts, it seems likely that there is an artefact of the data collection  
388 that affects predominantly the Persisters.

389

390 Terminations appear to be severely underreported for Persisters (Figure 6). Underreporting of  
391 terminations is potentially consequential for reproductive and medical studies (46). For  
392 example, biased termination records have led to erroneous reports of association between  
393 terminations and breast cancer, which were only later disproved through laborious large-scale  
394 studies (47,48). The underreporting is likely due to social-desirability bias. Specifically, there are  
395 two aspects of study design for the 1958 Birth Cohort that may have exacerbated it. The first is  
396 that most of the fertility data were elicited in interviews, whether face-to-face or by telephone,  
397 rather than in self-administered questionnaires which might mitigate social-desirability bias. The  
398 second is the change in fertility questionnaire structure between Sweep 4 (in 1981) and Sweep  
399 5 (in 1991). Whereas the participants in Sweep 4 (Age 23) were asked to list any natural children  
400 they may have, in birth order, from Sweep 5 (Age 33) onwards participants were asked

401 specifically about their pregnancies, starting from the most recent and going backward in  
402 time. Miscarriages and terminations from Sweep 4 (Age 23) are reported only collectively, as  
403 total numbers, without associated dates. National statistics (Figure 6) suggest that the  
404 terminations for women born in 1958 peaked by 1976. This means that the increased resolution  
405 in data collection for the cohort, introduced in Sweep 5 (Age 33), appears too late. The  
406 responses to the pregnancy questions from 1991 onwards, although informative in assisting to  
407 reconstruct the past, may be compromised by recall bias on top of the inevitable social-  
408 desirability bias, and the aforementioned non-random attrition.

409

410 For Persisters, we have found that participation in the final four sweeps (Biomedical (Age 44)  
411 and Sweeps 7-9 (Ages 46-55)) was associated with being older both at first birth and at  
412 departure from full-time education. That these two event times follow similar patterns might  
413 have been anticipated from positive correlation established in the literature (49,50), but we  
414 know of no reason to have predicted an association between the timings of those events and  
415 Continuation Length in the study.

416

417 The correlation between Continuation Length and the above-mentioned two event times  
418 becomes stronger when continuation is defined by a subject's engagement specifically with  
419 the fertility questions. How to measure Continuation Length is an important consideration,  
420 especially within the context of any statistical analyses of fertility – for instance, comparing  
421 menopause timing – that try to account for right censoring of missing subjects. Using Overall  
422 Continuation Length as the censoring time for individuals who lack event times may be  
423 misleading if the individuals have not interacted with any of the fertility questions.  
424 Fundamentally, the last time when we can confidently assume that an event has not occurred  
425 is the last time when we have been explicitly or implicitly told it has not occurred. When subjects  
426 have responded only to some questions we may be left unsure whether they are, in fact,  
427 censored. As in our case menopause timings are not captured by an external registry, the  
428 suggested correction proposed by Lesko et al. (51) does not apply.

429

430 Whether overall the 1958 British Birth Cohort is a good proxy for the British population of this age  
431 group – in other words, whether it is representative of the population – is a multifaceted  
432 question that defies any simple answer. The general definition of representativeness of cohorts  
433 in research on human populations, and whether or not it is an essential quality of a study, has  
434 been discussed extensively (25–30). The analyses presented here illustrate some of the ways in  
435 which representativeness may depend on the research question under consideration. This has  
436 implications for the use of the cohort study data. Once a research question has been defined,  
437 relevant biases can be established and, in some cases, quantified, for example by means of  
438 causal diagrams, relative odds ratios, or expert assessment (52,53). Then, depending on the  
439 combined nature of the bias and the research question, remedial action can be considered –  
440 whether it is a quantitative correction such as multiple imputation with inverse probability  
441 weighting (54,55), or by other means, such as those suggested in (56). Since it is *a priori* not  
442 known whether the effect of biases in associations are very modest, as in (52), or quite  
443 substantial, as in (54), it is important to determine how each specific research question may  
444 interact with the identified biases.

445 **ETHICS**

446 Access to anonymised linked records from NCDS was obtained after approval by the  
447 METADAC panel on April 11th 2017 (application number: MDAC-2017-0007-01).

448

449

450 **FUNDING**

451 This work was supported by the Biotechnology and Biological Sciences Research Council

452 [BB/S001824/1]

453 **ACKNOWLEDGMENTS**

454 Data governance was provided by the METADAC data access committee, funded by ESRC,  
455 Wellcome, and MRC. (Grant Number: MR/N01104X/1)

456 This work made use of data and samples generated by the 1958 Birth Cohort (NCDS), which is  
457 managed by the Centre for Longitudinal Studies at the UCL Institute of Education, funded by  
458 the Economic and Social Research Council (grant number ES/M001660/1). Access to these  
459 resources was enabled via the Wellcome Trust & MRC: 58FORWARDS grant [108439/Z/15/Z] (The  
460 1958 Birth Cohort: Fostering new Opportunities for Research via Wider Access to Reliable Data  
461 and Samples). Before 2015 biomedical resources were maintained under the Wellcome Trust  
462 and Medical Research Council 58READIE Project (grant numbers WT095219MA and G1001799).



463 **REFERENCES**

- 464 1. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development  
465 Study). *Int J Epidemiol*. 2006 Feb 1;35(1):34–41.
- 466 2. Atherton K, Fuller E, Shepherd P, Strachan DP, Power C. Loss and representativeness in a  
467 biomedical survey at age 45 years: 1958 British birth cohort. *J Epidemiol Community*  
468 *Health*. 2008;62(3):216–23.
- 469 3. Capon F, Di Meglio P, Szaub J, Prescott NJ, Dunster C, Baumber L, et al. Sequence  
470 variants in the genes for the interleukin-23 receptor (IL23R) and its ligand (IL12B) confer  
471 protection against psoriasis. *Hum Genet*. 2007;122(2):201–6.
- 472 4. Simpson CL, Hysi P, Bhattacharya SS, Hammond CJ, Webster A, Peckham CS, et al. The  
473 roles of PAX6 and SOX2 in myopia: Lessons from the 1958 British Birth Cohort. *Investig*  
474 *Ophthalmol Vis Sci*. 2007;48(10):4421–5.
- 475 5. Capon F, Bijlmakers M-J, Wolf N, Quaranta M, Huffmeier U, Allen M, et al. Identification  
476 of ZNF313 / RNF114 as a novel psoriasis susceptibility gene. *Hum Mol Genet*. 2008 Jul  
477 1;17(13):1938–45.
- 478 6. Razavi T, Clark C, Stansfeld SA. Work-family conflict as a predictor of common mental  
479 disorders in the 1958 British birth cohort. *Longit Life Course Stud*. 2015;6(3):264–78.
- 480 7. Kwok MK, Yeung SLA, Leung GM, Mary Schooling C. Birth weight and adult  
481 cardiovascular risk factors using multiple birth status as an instrumental variable in the  
482 1958 British Birth Cohort. *Prev Med (Baltim)*. 2016;84:69–75.
- 483 8. Staff J, Maggs J. Alcohol and Cigarette Use From Ages 23 to 55: Links With Health and  
484 Well-Being in the Long-Term National Child Development Study. *J Stud Alcohol Drugs*.  
485 2017;78(3):394–403.
- 486 9. Evans-Lacko S, Takizawa R, Brimblecombe N, King D, Knapp M, Maughan B, et al.  
487 Childhood bullying victimization is associated with use of mental health services over five  
488 decades: A longitudinal nationally representative cohort study. *Psychol Med*.  
489 2017;47(1):127–35.

- 490 10. Cooper R, Atherton K, Power C. Gestational age and risk factors for cardiovascular  
491 disease: Evidence from the 1958 British birth cohort followed to mid-life. *Int J Epidemiol.*  
492 2009;38(1):235–44.
- 493 11. Vimalleswaran KS, Cavadino A, Hyppönen E. Evidence for a genetic interaction in  
494 allergy-related responsiveness to vitamin D deficiency. *Allergy.* 2012 Aug;67(8):1033–40.
- 495 12. Vimalleswaran KS, Cavadino A, Hyppönen E. APOA5 genotype influences the  
496 association between 25-hydroxyvitamin D and high density lipoprotein cholesterol.  
497 *Atherosclerosis.* 2013 May;228(1):188–92.
- 498 13. Pinto Pereira SM, Li L, Power C. Early-life predictors of leisure-time physical inactivity in  
499 midadulthood: Findings from a prospective british birth cohort. *Am J Epidemiol.*  
500 2014;180(11):1098–108.
- 501 14. Krzyzanowska M, Mascie-Taylor CGN, Thalabard JC. Is human mating for height  
502 associated with fertility? Results from a british national cohort study. *Am J Hum Biol.*  
503 2015;27(4):553–63.
- 504 15. Lacey RE, Sacker A, Kumari M, Worts D, McDonough P, Booker C, et al. Work-family life  
505 courses and markers of stress and inflammation in mid-life: Evidence from the National  
506 Child Development Study. *Int J Epidemiol.* 2016;45(4):1247–59.
- 507 16. Cherrie MPC, Sarran C, Osborne NJ. Association between serum 25-hydroxy vitamin D  
508 levels and the prevalence of adult-onset asthma. *Int J Environ Res Public Health.*  
509 2018;15(6):1–12.
- 510 17. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al.  
511 Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity  
512 variants. *Nat Genet.* 2007;39(11):1329–37.
- 513 18. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al.  
514 Genome-wide association study of 14,000 cases of seven common diseases and 3,000  
515 shared controls. *Nature.* 2007;447(7145):661–78.
- 516 19. Speedy HE, Di Bernardo MC, Sava GP, Dyer MJS, Holroyd A, Wang Y, et al. A genome-  
517 wide association study identifies multiple susceptibility loci for chronic lymphocytic

- 518 leukemia. *Nat Genet.* 2014;46(1):56–60.
- 519 20. Law PJ, Sud A, Mitchell JS, Henrion M, Orlando G, Lenive O, et al. Genome-wide  
520 association analysis of chronic lymphocytic leukaemia, Hodgkin lymphoma and multiple  
521 myeloma identifies pleiotropic risk loci. *Sci Rep.* 2017;7(January):1–11.
- 522 21. Richards AL, Leonenko G, Walters JT, Kavanagh DH, Rees EG, Evans A, et al. Exome  
523 arrays capture polygenic rare variant contributions to schizophrenia. *Hum Mol Genet.*  
524 2016 Mar 1;25(5):1001–7.
- 525 22. Vignal CM, Bansal AT, Balding DJ. Using Penalised Logistic Regression to Fine Map HLA  
526 Variants for Rheumatoid Arthritis. *Ann Hum Genet.* 2011 Nov;75(6):655–64.
- 527 23. Bukodi E, Goldthorpe JH. Educational attainment - relative or absolute - as a mediator  
528 of intergenerational class mobility in Britain. *Res Soc Stratif Mobil.* 2016;43:5–15.
- 529 24. Bukodi E, Goldthorpe JH, Waller L, Kuha J. The mobility problem in Britain: New findings  
530 from the analysis of birth cohort data. *Br J Sociol.* 2015;66(1):93–117.
- 531 25. Rothman KJ, Gallacher JEJ, Hatch EE. Why representativeness should be avoided. *Int J*  
532 *Epidemiol.* 2013;42(4):1012–4.
- 533 26. Elwood JM. Commentary: On representativeness. *Int J Epidemiol.* 2013;42(4):1014–5.
- 534 27. Nohr EA, Olsen J. Commentary: Epidemiologists have debated representativeness for  
535 more than 40 years-has the time come to move on? *Int J Epidemiol.* 2013;42(4):1016–7.
- 536 28. Richiardi L, Pizzi C, Pearce N. Commentary: Representativeness is usually not necessary  
537 and often should be avoided. *Int J Epidemiol.* 2013;42(4):1018–22.
- 538 29. Ebrahim S, Smith GD. Commentary: Should we always deliberately be non-  
539 representative? *Int J Epidemiol.* 2013;42(4):1022–6.
- 540 30. Rothman KJ, Gallacher JEJ, Hatch EE. Rebuttal: When it comes to scientific inference,  
541 sometimes a cigar is just a cigar. *Int J Epidemiol.* 2013;42(4):1026–8.
- 542 31. Grépin KA, Bharadwaj P. Maternal education and child mortality in Zimbabwe. *J Health*  
543 *Econ.* 2015;44:97–117.
- 544 32. Mills M, Rindfuss RR, McDonald P, te Velde E. Why do people postpone parenthood?  
545 Reasons and social policy incentives. *Hum Reprod Update.* 2011;17(6):848–60.

- 546 33. Monstad K, Propper C, Salvanes KG. Education and fertility: Evidence from a natural  
547 experiment. *Scand J Econ*. 2008;110(4):827–52.
- 548 34. Kravdal Ø, Rindfuss RR. Changing relationships between education and fertility: A study  
549 of women and men born 1940 to 1964. *Am Sociol Rev*. 2008;73(5):854–73.
- 550 35. National Records of Scotland. Births Time Series Data - Births, by mother's age, Scotland,  
551 1945 to 2018. 2019. p. BT2.
- 552 36. Office for National Statistics. Dataset: Conception statistics. 1985.
- 553 37. Office for National Statistics. Dataset: Conception statistics [Internet]. 2003 [cited 2018  
554 Jan 22].
- 555 38. University of California Berkeley (USA), Max Planck Institute for Demographic Research  
556 (Germany). Human Mortality Database [Internet]. [cited 2018 Aug 18].
- 557 39. R Core Team. R: A language and environment for statistical computing. R Foundation  
558 for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. [Internet].  
559 Vienna, Austria.: R Foundation for Statistical Computing; 2019.
- 560 40. Wickham H. tidyverse: Easily Install and Load the “Tidyverse”. [Internet]. 2017.
- 561 41. Therneau T. A Package for Survival Analysis in S version 2.38 URL: [https://CRAN.R-](https://CRAN.R-project.org/package=survival)  
562 [project.org/package=survival](https://CRAN.R-project.org/package=survival). 2015.
- 563 42. R Studio Team. RStudio: Integrated Development for R [Internet]. Boston, MA; 2018.
- 564 43. Grigoriev P, Jdanov DA, Shkolnikov VM. New methods for estimating detailed fertility  
565 schedules from abridged data. 2018;49(0):0–27.
- 566 44. Brilleman SL, Pachana NA, Dobson AJ. The impact of attrition on the representativeness  
567 of cohort studies of older people. *BMC Med Res Methodol*. 2010;10.
- 568 45. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The  
569 Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)  
570 statement: Guidelines for reporting observational studies. *Epidemiology*. 2007;18(6):800–  
571 4.
- 572 46. Fu H, Darroch JE, Henshaw SK, Kolb E. Measuring the extent of abortion underreporting  
573 in the 1995 National Survey of Family Growth. *Fam Plann Perspect*. 1995;30(3):128–33,

- 574 138.
- 575 47. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and abortion:  
576 Collaborative reanalysis of data from 53 epidemiological studies, including 83 000  
577 women with breast cancer from 16 countries. *Lancet*. 2004;363(9414):1007–16.
- 578 48. Stuart GS, Grimes DA. Social desirability bias in family planning studies: a neglected  
579 problem. *Contraception*. 2009;80(2):108–12.
- 580 49. Rendall M, Aracil E, Bagavos C, Couet C, DeRose A, DiGiulio P, et al. Increasingly  
581 heterogeneous ages at first birth by education in Southern European and Anglo-  
582 American family-policy regimes: A seven-country comparison by birth cohort. *Popul*  
583 *Stud (NY)*. 2010;64(3):209–27.
- 584 50. Bongaarts J, Mensch BS, Blanc AK. Trends in the age at reproductive transitions in the  
585 developing world: The role of education. *Popul Stud (NY)*. 2017;71(2):139–54.
- 586 51. Lesko CR, Edwards JK, Cole SR, Moore RD, Lau B. When to Censor? *Am J Epidemiol*.  
587 2018;187(3):623–32.
- 588 52. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. *Acta*  
589 *Obstet Gynecol Scand*. 2018;97(4):407–16.
- 590 53. Thompson S, Ekelund U, Jebb S, Lindroos AK, Mander A, Sharp S, et al. A proposed  
591 method of bias adjustment for meta-analyses of published observational studies. *Int J*  
592 *Epidemiol*. 2011;40(3):765–77.
- 593 54. Munafò MR, Tilling K, Taylor AE, Evans DM, Smith GD. Collider scope: When selection bias  
594 can substantially influence observed associations. *Int J Epidemiol*. 2018;47(1):226–35.
- 595 55. Mostafa T, Wiggins RD. The impact of attrition and non-response in birth cohort studies:  
596 A need to incorporate missingness strategies. *Longit Life Course Stud*. 2015;6(2):131–46.
- 597 56. Ioannidis JPA. Commentary: Adjusting for bias: A user's guide to performing plastic  
598 surgery on meta-analyses of observational studies. *Int J Epidemiol*. 2011;40(3):777–9.
- 599