

Cognitive development following ART: effect of choice of comparison group, confounding and mediating factors

C. Carson^{1,4}, J.J. Kurinczuk¹, A. Sacker², Y. Kelly³, R. Klemetti¹, M. Redshaw¹, and M.A. Quigley¹

¹National Perinatal Epidemiology Unit, University of Oxford, Headington, Oxford OX3 7LF, UK ²Institute of Social and Economic Research, University of Essex, Colchester, Essex CO4 3SQ, UK ³Department of Epidemiology & Public Health, University College London, Gower Street Campus, London WC1E 6BT, UK

⁴Correspondence address. E-mail: claire.carson@npeu.ox.ac.uk

BACKGROUND: Epidemiological studies have examined the health of children born after assisted reproductive technology (ART), with contradictory results. In this article, we address the question ‘Do singletons born after ART have a poorer cognitive developmental outcome at 3 years of age?’ We assess the implications of using different comparison groups, and discuss appropriate analytical approaches for the control of confounding and mediating variables.

METHODS: Data were drawn from the Millennium Cohort Study. Interviews captured sociodemographic, behavioural and pregnancy information. Developmental assessments conducted at age three included the British Ability Scales II Naming Vocabulary (BAS-NV) instrument. We compared ART infants (born after IVF or ICSI) to four comparison groups: a ‘matched’ group; a ‘subfertile’ group (time to conception > 12 months); a ‘fertile’ group (time to conception < 12 months); and an ‘any spontaneous conceptions’ group. Linear regression provided estimates of the difference in mean BAS-NV scores in the ART and comparison groups; both unadjusted estimates and those adjusted for confounding and mediating factors are presented.

RESULTS: In the unadjusted analyses, ART children gained significantly better BAS-NV test results than did the comparison group children. When converted to an estimate of developmental age gap, ART children were 2.5, 2.7, 3.6 and 4.5 months ahead of the ‘matched’, ‘subfertile’, ‘fertile’ and ‘spontaneous conception’ children, respectively. After adjusting for confounding and mediating factors, the differences were reduced, and were not statistically significant.

CONCLUSIONS: ART is not associated with poorer cognitive development at 3 years. We have highlighted methodological considerations for researchers planning to study the effect of infertility and ART on childhood outcomes.

Key words: ART / comparison groups / methodology / confounding factors / ART children

Introduction

The use of assisted reproductive technology (ART) is becoming more commonplace as the number of people seeking treatment rises (Oakley *et al.*, 2008). The Human Fertilization and Embryology Authority recorded 44 275 treatment cycles in the UK in 2006, resulting in 12 596 ART children (an increase of 11.9% since 2005). Although such success gives hope to infertile couples, the techniques used have become increasingly invasive, and there is concern about the health and development of children born following IVF and ICSI.

There is good evidence that preterm birth, lower birthweight (Helmerhorst *et al.*, 2004; Klemetti *et al.*, 2006) and congenital

malformations (Hansen *et al.*, 2002, 2008) are more common in singletons born after IVF, but the data for other outcomes are less clear. Epidemiological studies have examined the physical, neurological and developmental health of children born after ART, with contradictory results. There is some suggestion that IVF children are more likely to have suffered at least one childhood illness before the age of three (Koivurova *et al.*, 2003), and experience more hospitalizations by 6 years of age (Ericson *et al.*, 2002; Klemetti *et al.*, 2006). Most research has found no difference between IVF children and spontaneously conceived children in terms of vision, hearing or growth (Saunders *et al.*, 1996; Pinborg *et al.*, 2003; Place and Englert 2003), though a single larger study has found higher rates of severe visual problems in IVF

children (Stromberg *et al.*, 2002). Again, the evidence relating to developmental outcomes is equivocal (Hvidtjorn *et al.*, 2009); most small studies report no effect whereas the larger studies have found an increased risk of developmental delay (Bowen *et al.*, 1998; Stromberg *et al.*, 2002; Sutcliffe *et al.*, 2003; Agarwal *et al.*, 2005; Lidegaard *et al.*, 2005; Leunens *et al.*, 2006; Knoester *et al.*, 2008). The evidence for adverse effects on neurological health is stronger, with sizeable cohort studies reporting increased risks of both cerebral palsy (Ericson *et al.*, 2002; Stromberg *et al.*, 2002; Lidegaard *et al.*, 2005; Hvidtjorn *et al.*, 2009) and epilepsy (Ericson *et al.*, 2002), though how much of this is mediated through the effect of preterm birth is unclear (Middelburg *et al.*, 2008).

The studies that produced these findings used a variety of designs and comparison groups. Some exploit existing data, using registry data to compare children born after ART with the rest of the non-ART population (Doyle *et al.*, 1998; Kallen *et al.*, 2005; Lidegaard *et al.*, 2005). Others make use of a convenient population, such as couples attending fertility clinics (Klip *et al.*, 2001; Ceelen *et al.*, 2008) or maternity hospitals (Wennerholm *et al.*, 1998). Another approach is to recruit infants conceived using ART and then select a comparison group matched on potential confounding factors such as gestational age (Sutcliffe *et al.*, 1995, 2003; Koudstaal *et al.*, 2000; Stromberg *et al.*, 2002). The choice of comparison group will affect the interpretation of the results, while the source of the study population can restrict the investigation of important confounding or mediating factors. A recent systematic review noted that many existing studies of the neurodevelopmental effects of ART were difficult to interpret due to methodological shortcomings (Middelburg *et al.*, 2008).

In this paper we assess the implications of using different comparison groups in our analyses, and discuss appropriate analytical approaches for the control of confounding and mediating variables. To do this, we address the question 'Do singletons born after ART have a poorer cognitive developmental outcome at three years of age?'

Materials and Methods

The data were drawn from the Millennium Cohort Study (MCS). Details of the MCS methods have been described in detail elsewhere (Dex and Joshi, 2005; Plewis, 2007a, b). Briefly, a random two stage sample of all infants (both singletons and multiples) born in England and Wales between September 2000 and August 2001, and in Scotland and Northern Ireland between November 2000 and January 2002, and who were resident in the UK at 9 months, was drawn from Department of Social Security Child Benefit Registers (to which every child is automatically entitled). Ethnically diverse and disadvantaged wards were over-sampled to ensure adequate representation of such areas. Baseline interviews captured sociodemographic and health information, including pregnancy details, from 18 553 families (comprising 246 sets of twins, 10 sets of triplets and 18 297 singletons thus totalling 18 819 children). Of the mothers, 54% (9979/18 553) reported a planned pregnancy, and were asked about time to conception and assisted reproductive treatments. A total of 80% (14 898/18 553) completed follow-up questionnaires and assessments when the children were 3 years old (mean 3.1 years, range 2.7, 4.6). Cognitive development was assessed, using the Naming Vocabulary instrument from the British Ability Scales II (BAS-NV), which measures expressive language ability (Elliott *et al.*, 1997).

To illustrate the effect of selecting and using different comparison groups on the association between ART and subsequent health outcomes, four analyses were conducted. The 'exposed' group of ART infants include those children whose mother's reported a birth resulting from IVF (including frozen embryo transfer, FET) or ICSI. Four separate comparison groups were identified:

- (i) A matched comparison group (MC) was drawn from the MCS dataset with a ratio of two comparison children to each ART case, matching on gender, maternal age (in years) and socioeconomic status (NS-SEC, based on occupation).
- (ii) The second group comprised children born to subfertile parents who did not undergo fertility treatment. These children were born following planned pregnancies, with a prolonged time to conception (PTTC) defined as >12 months. This comparison group resembles that often used in studies based in infertility clinics.
- (iii) The third comparison group was drawn from planned pregnancies with a normal time to conception (NTTC) defined as <12 months. These are children born to 'fertile' couples.
- (iv) Finally, a spontaneous conception comparison group (SC) included all singletons recruited in the MCS who were not the product of IVF or ICSI. This comparison group contains fertile couples (unplanned pregnancies or planned pregnancies with a normal time to conception) and subfertile couples (some of whom may have received non-ART infertility treatment). All children in the first three comparison groups are included in this fourth, least selective, comparison group.

The analyses using each of the comparison groups (MC, PTTC, NTTC and SC) were completed as follows: maternal and infant characteristics were described in each group; linear regression provided estimates of the difference in mean BAS-NV scores in the ART and comparison groups; both crude estimates and those adjusted for confounding and mediating factors are presented. Variables were considered confounding factors if they were significantly ($P < 0.05$) associated with both the exposure (ART) and the outcome (BAS Naming Vocabulary score) and altered the odds ratio by more than 10%. Maternal age, socioeconomic status at baseline and baby's sex were included on an *a priori* basis. Other variables considered as potential confounders were mother's marital status, income and educational attainment at baseline, parity, smoking or drinking in pregnancy, breastfeeding, and the age of the child when the BAS test was completed. The role of potential mediating factors in explaining any observed relationships was also investigated—these variables could be on the causal pathway between ART and later child development: preterm birth, low birthweight and differences in parental involvement at 3 years of age indicated by the Pianta Parent–Child Relationship score (Pianta, 1995), frequency of reading to the child and reported hours of television the child watched per day.

To aid comparability between the four comparison groups, models were adjusted for the same covariates. Model 1 shows the crude association between ART and BAS-NV, using each comparison group. Model 2 controls only for the variables that are most often available to researchers (maternal age, socioeconomic status, baby's sex, and age of the child when BAS-NV was conducted). Model 3 additionally controls for those variables considered confounding factors but for which data are less commonly available (parity and alcohol in pregnancy). Model 4 also adjusts for mediating variables that occur in early life (gestational age and birthweight). The final Model 5 is also adjusted for mediating variables, which occur in later childhood (i.e. differences in parenting). To aid interpretation, the differences in mean BAS-NV score are also presented as 'months of developmental delay' (Elliott *et al.*, 1997).

This analysis was limited to singletons, for whom data were available on pregnancy, cognitive outcome at 3 years and key confounding factors ($N =$

10 583). All analyses were conducted in STATA 10 (StataCorp, 2007), using survey commands to allow for the clustered design, with weights taking into account over-sampling at sweep 1 and non-response at sweep 2 (Plewis, 2007a, b).

Results

The characteristics of ART and comparison groups

Table I shows clear differences between the ART group and the comparison groups, and also between the comparison groups themselves. Although 71% of the ART mothers reported that this was their first birth, the proportion in the comparison groups was much lower (35% for MC, 50% for PTTC 40% for NTTC and 43% for SC). ART mothers were significantly older than the PTTC, NTTC or SC

mothers (mean 35 years, compared with 32, 31 and 30 years, respectively), and were more likely: to be married, to be of a more advantageous socioeconomic status, to have a university degree and to report a higher household income. Compared with the ART mothers, significantly more women in the PTTC, NTTC and SC groups reported consuming alcohol during their pregnancies. Although the proportion of smokers in all the comparison groups was higher than in the ART mothers, the difference was only significant for the MC group.

All comparison groups included children that were, on average, born at a later gestational age than the ART group. The proportion of ART children born preterm (<37 weeks) was nearly three times that seen in the SC group (18.9 versus 6.5%). Similarly, the proportion of low birthweight children (<2.5 kg) in the ART group was significantly higher than in any comparison group (13.8% in ART, compared with 5.0% in MC, 6.2% in PTTC, 4.3% in NTTC and 5.3% SC). The

Table I Characteristics of the ART group, and each possible comparison group: individuals with full data only

Characteristic	ART singletons (ICSI, IVF, FET)	Matched group (MC) [£]	Prolonged time to pregnancy (PTTC)	Normal time to pregnancy (NTTC)	Any spontaneous conceptions (SC)
N (unweighted)	99	198	402	5556	10 574
Maternal characteristics (at sweep 1)					
This baby is her first birth (%)	71.3	35.3*	50.3*	39.8*	42.6*
Age, years (mean)	35.1	35.3	32.1*	30.8*	29.7*
Married (%)	89.7	73.4	78.7	76.0*	60.7*
Manual socioeconomic status ⁺ (%)	13.2	8.0	22.3	21.5	31.6*
Income < £10 400 per year [§] (%)	4.5	4.3	9.9	9.0	19.5*
University degree (%)	52.1	57.3	36.6	44.4	36.2*
Smoked while pregnant (%)	3.8	9.8	16.9*	13.5*	20.9*
Drank alcohol in pregnancy (%)	27.9	46.0*	30.4	37.4	35.2
Infant characteristics					
Gestational age, weeks (mean)	38.5	39.1	39.2*	39.4*	39.3*
Preterm birth, <37 weeks (%)	18.9	6.4*	7.5*	5.5*	6.5*
Birthweight, g (mean)	3204	3398*	3361	3449*	3405*
Low birthweight, <2.5 kg (%)	13.8	5.0*	6.2*	4.3*	5.3*
Male sex (%)	47.0	42.4	50.9	50.7	50.3
Breastfed at all for ≥4 months (%)	39.9	45.8	29.6	39.5	34.0
Age at Sweep 2 interview, years (mean)	3.11	3.10	3.13	3.12	3.12
Parenting variables					
Pianta parent-child relationship Inventory score (mean, lower number indicates poorer relationship)	64.8	65.5	64.5	64.9	64.4
Reads to child every day (%)	81.8	68.6*	68.5*	66.9*	62.8*
Child watches >3 h television each day (%)	12.8	10.3	15.5	12.4	12.5
Main outcome of interest					
British ability scales, naming vocabulary at 3 years (mean ability score)	81.6	78.5	78.3*	77.1*	76.0*

Means and proportions presented are weighted to account for clustering, stratification and non-response at sweep2.

[£]Matched on: maternal age, social class and baby's sex.

*Significantly different from the ART group $P < 0.05$ for design based χ^2 tests for proportions and t tests for difference in means.

⁺Highest SEC of either parent is 'routine and manual' or 'never worked or long-term unemployed'.

[§]Combined income of both parents in two-parent families.

observed difference in mean birthweight at the population level of 200 g (ART children compared with all other births) is considerable, and equivalent to the effect of cigarette smoking (British Medical Association, 2004).

Indicator variables for parenting behaviour differed between the groups. Though there was no indication that the quality of the maternal child relationship varied across the groups, the variable indicating parental involvement (daily reading to the child) showed significant differences with 82% of ART parents reporting this activity, compared with just 63% in the SC group.

The effect of ART on subsequent BAS-NV at 3 years, and differences between the four analyses

In general the observed effects of adjustment for confounders on BAS-NV were similar with each comparison group (Tables II and III). In the unadjusted analyses, ART children achieved significantly better scores in the BAS-NV tests than the comparison group children (Model 1). In general, after adjusting for confounding factors, the effect was reduced, and the difference was not statistically significant (Models 2 and 3). Further adjustment for mediating factors early in the lifecourse inflates the difference between the ART and comparison

group children (Model 4), whereas adjusting for the later mediating influence of parental involvement reduces the difference (Model 5).

When one looks in more detail at the results, differences are observed between the four different models (Fig. 1). In the crude analyses a statistically significant increase was observed in the mean BAS-NV scores for ART children compared with the PTTC, NTTC and SC groups. The positive effect of ART on child development at 3 years appeared greater when compared with groups that were least like the ART group. That is, when compared with the matched group (who would be considered most similar to the ART group in terms of behaviour and characteristics) the difference was 3.12 points, which increased to 3.39 when compared with the subfertile PTTC group, then to 4.52 in the fertile NTTC group and finally 5.67 in the SC group (who would be considered least like the ART group). When converted to an estimate of developmental age gap, ART children were 2.5, 2.7, 3.6 and 4.5 months *ahead* of MC, PTTC, NTTC and SC children, respectively.

Adjusting for *a priori* confounders (maternal age, socioeconomic status and child's age at testing) in the PTTC, NTTC and SC analyses, reduced the difference in means. In the PTTC group, a significant difference was no longer seen.

Additionally, controlling for the confounding effect of alcohol in pregnancy and whether this child was the mother's first birth, ameliorated the effect of ART further, so that there was no significant

Table II The difference in the mean BAS-NV[#] ability score^{**} between the ART group and each comparison group, for the four analytical models

Comparison group	Model 1	Model 2	Model 3	Model 4	Model 5
Matched (MC)	3.12 (−0.71, 6.94)	3.94 (0.49, 7.39)*	1.61 (−2.02, 5.24)	1.76 (−1.87, 5.39)	1.10 (−2.47, 4.68)
Prolonged (PTTC)	3.39 (0.01, 6.76)*	1.49 (−1.67, 4.64)	0.82 (−2.33, 3.96)	1.42 (−1.76, 4.60)	1.44 (−1.74, 4.63)
Normal (NTTC)	4.52 (1.44, 7.61)*	2.94 (0.20, 5.67)*	1.15 (−1.81, 4.11)	1.74 (−1.28, 4.75)	1.66 (−1.34, 4.66)
Spontaneous (SC)	5.67 (2.61, 8.72)*	3.56 (0.89, 6.27)*	1.41 (−1.48, 4.311)	1.93 (−1.01, 4.88)	1.84 (−1.07, 4.74)

Model 1: Crude association. Note that the result for the matched analysis is not truly a crude estimate, as groups are matched on maternal age, social class and child's sex.

Model 2: As Model 1, but also adjusted for available data—maternal age, social class, child's sex, child's age at assessment.

Model 3: As Model 2, but also adjusted for true confounding factors—parity, alcohol in pregnancy.

Model 4: As Model 3, but also adjusted for mediating variables early in the life course—gestational age (weeks) and birthweight (kg).

Model 5: As Model 4, but also adjusted for mediating variables later in the life course—maternal–child relationship and frequency reading to the child.

[#]BAS-NV: British Ability Scales II Naming Vocabulary instrument.

^{**}These figures represent the difference (coefficient) in mean BAS-NV score between the comparison group and the ART group—positive numbers indicate that the ART children have a higher mean score than the comparison children.

*Indicates a significant difference in mean BAS score, $P < 0.05$ between the comparison group and the ART group.

Table III The difference in BAS-NV[#] ability score between the ART group and each comparison group, converted into the equivalent developmental age gap (in months)

Comparison group	Model 1	Model 2	Model 3	Model 4	Model 5
Matched (MC)	2.5	3.2	1.3	1.4	0.9
Prolonged (PTTC)	2.7	1.2	0.7	1.1	1.2
Normal (NTTC)	3.6	2.4	0.9	1.4	1.3
Spontaneous (SC)	4.5	2.8	1.1	1.5	1.5

Age equivalent in months is based on an estimated difference in BAS-NV ability score of 1.25 [from MCS Guide to the datasets (Hansen, 2008)]. These figures represent the number of months that the ART children are *ahead* of the comparison group.

[#]BAS-NV: British Ability Scales II Naming Vocabulary instrument.

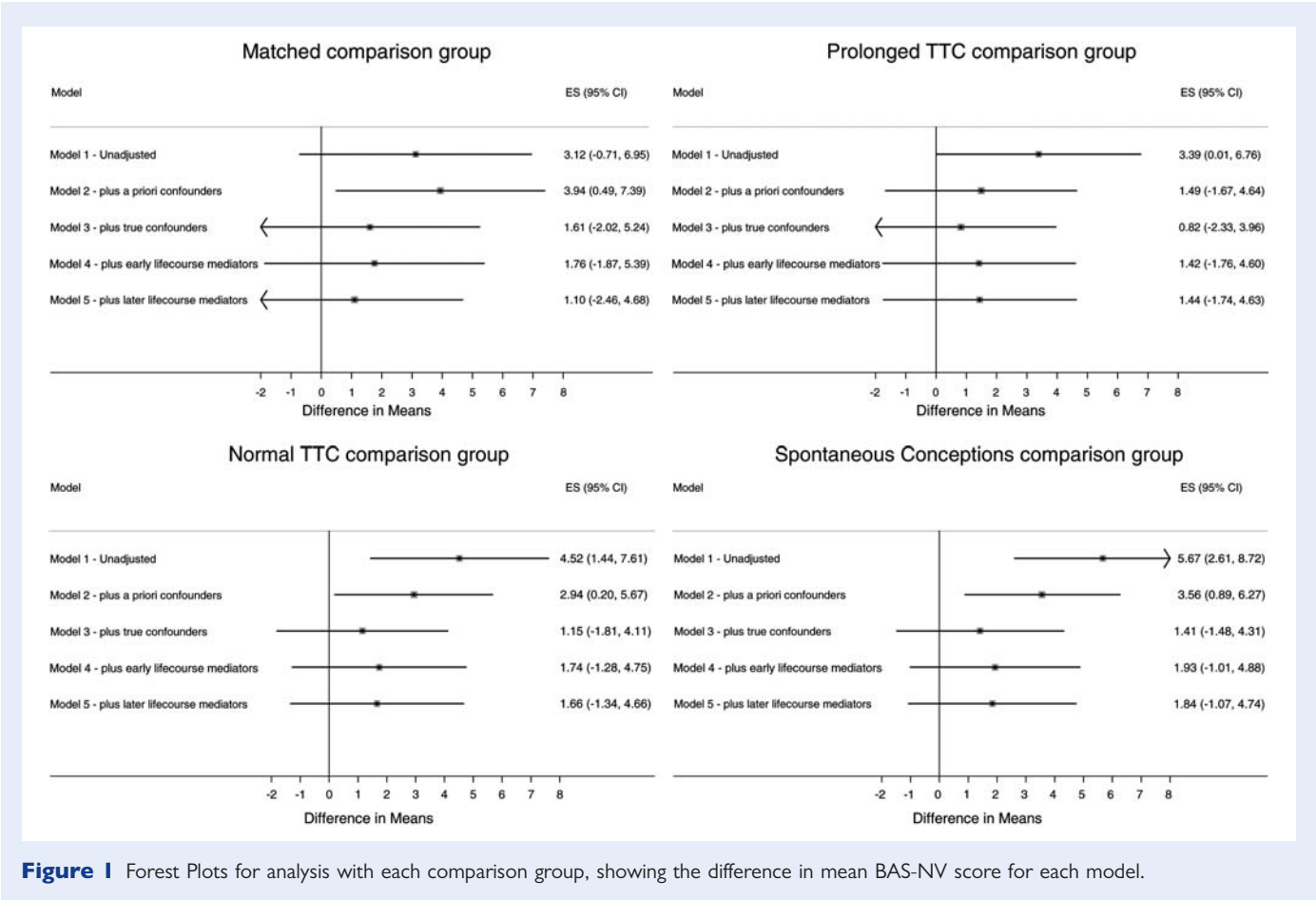


Figure 1 Forest Plots for analysis with each comparison group, showing the difference in mean BAS-NV score for each model.

difference between the ART and any comparison group. However, the estimates of difference in mean BAS-NV score continued to differ; the matched analysis suggested that ART children were 1.3 months ahead, whereas the PTTC analysis suggested there was just less than 1 month advantage.

Adjusting for mediating variables increased the difference between the ART and comparison groups. Since ART children were more likely to be born earlier and at lower birthweight and these factors are associated with poorer developmental outcomes, adjusting for these mediating variables increased the observed difference in mean BAS-NV score (Model 4), whereas controlling for the effects of increased parental involvement slightly reduced the effect of ART in the NTTC and MC analyses (Model 5).

Discussion

The effect of art on subsequent child development scores at 3 years

Research has demonstrated that children born following ART are at greater risk of some adverse health outcomes, but there are few good epidemiological studies of neurodevelopmental effects (Middelburg et al., 2008). We investigated whether ART children were disadvantaged in terms of cognitive development at 3 years of age as indicated by the BAS Naming Vocabulary test scores. Our

findings show that children born after ART appear to have better expressive language abilities than children born after non-ART conceptions. In general, women who undergo ART tend to be older, more highly educated, and socially and economically advantaged. Consequently, we would expect children born to such women to have advantages that may improve their attainment in cognitive tests. Leunens et al. (2006) reported higher intelligence test scores in ICSI children, which disappears after controlling for the effect of maternal education. Our analyses confirmed this, and after adjusting for confounding factors (particularly maternal age, socioeconomic status, number of children and alcohol in pregnancy), the significantly higher BAS-NV scores among ART children disappear.

Assisted reproductive technologies are associated with a range of other risks and behaviours that may be on the causal pathway between ART and neurodevelopmental outcomes. Some are biological effects, such as an increased risk of preterm birth and low birthweight, which could result in poorer cognitive development (Agarwal et al., 2005; Klemetti et al., 2006). Adjusting for these potential mediating factors increased the apparent advantage that ART children have over non-ART children suggesting that even given the disadvantage of poorer birth outcomes, ART children did not on average suffer adverse developmental effects although the overall effect remained statistically non-significant.

Contrasting experiences of conception and pregnancy may result in behavioural differences between ART and non-ART parents. ART

children are usually conceived after a prolonged period of attempted conception, a diagnosis of infertility and invasive (and potentially costly) medical treatment. For these reasons, ART patients may consider their children to be more 'precious', and may (consciously or unconsciously) invest more in their parenting (Golombok *et al.*, 1995). ART mothers may also have greater social capital and 'intellectual resources' at their disposal, for example 58% of the mothers reported a degree qualification, which may alter their parenting behaviour. We found that whereas the reported quality of the maternal-child relationship did not differ between ART and non-ART families, indicators of parenting behaviour did vary. More ART mothers reported reading to their child each day, which is consistent with research that shows greater commitment to parenting and greater parent-child interaction in ART families (Barnes *et al.*, 2004). Reading stories and picture books helps children to increase their vocabulary and may offer an advantage in the BAS-NV tests, thus conferring an advantage on ART children. This was confirmed by controlling for the effects of these behavioural factors which slightly reduced the difference between ART and non-ART children.

Methodological issues

Choosing an appropriate comparison group for your research question

Though the overarching pattern of results is the same for each analysis, we demonstrated some important differences depending upon the choice of comparison group. Although the findings differ, it is important to note that none should be regarded as 'wrong'. Sociodemographic and behavioural factors vary between the groups, showing that the comparison groups are not surprisingly capturing different parts of the general population. Thus the four analyses address

different research questions, and therefore the estimates of effect apply to different hypotheses.

Clear identification of the research question is critical, as this guides the choice of an appropriate comparison group. Patients who undergo ART are, by definition, subfertile. By choosing the correct comparison group it is possible to investigate the effect of ART over and above any effect of infertility, or to look at the combined effect of infertility and ART on childhood outcomes (Buck Louis *et al.*, 2005). Figure 2 summarizes some key points when choosing a comparison group.

The effect of ART alone (ignoring any additional adverse effect of infertility) can only be assessed by comparing ART children to children born to subfertile parents who conceived without ART. Our ART/PTTC analysis takes this approach, comparing subfertile and infertile women in the general population. This is similar to studies conducted among infertility clinic patients where the comparison group is often spontaneous conceptions in couples awaiting treatment, although it is important to note that not all infertile women in the general population will attend a clinic for treatment. In the unadjusted model, we find that ART children have significantly better BAS-NV scores compared with this group but once we control for the effect of confounding factors the remaining difference is no longer significant.

The combined effect of infertility and its treatment perhaps has greater application to the planning and prediction of effects in the general population. To examine this, it is necessary to compare ART children with children born to couples without either infertility or ART. There are numerous ways to achieve this. The first is to compare ART infants with any spontaneous conception (i.e. all infants conceived without ART). The great advantage of this group is that the infants are easily identifiable and plentiful. Using existing resources such as birth registry data, can provide data on a huge

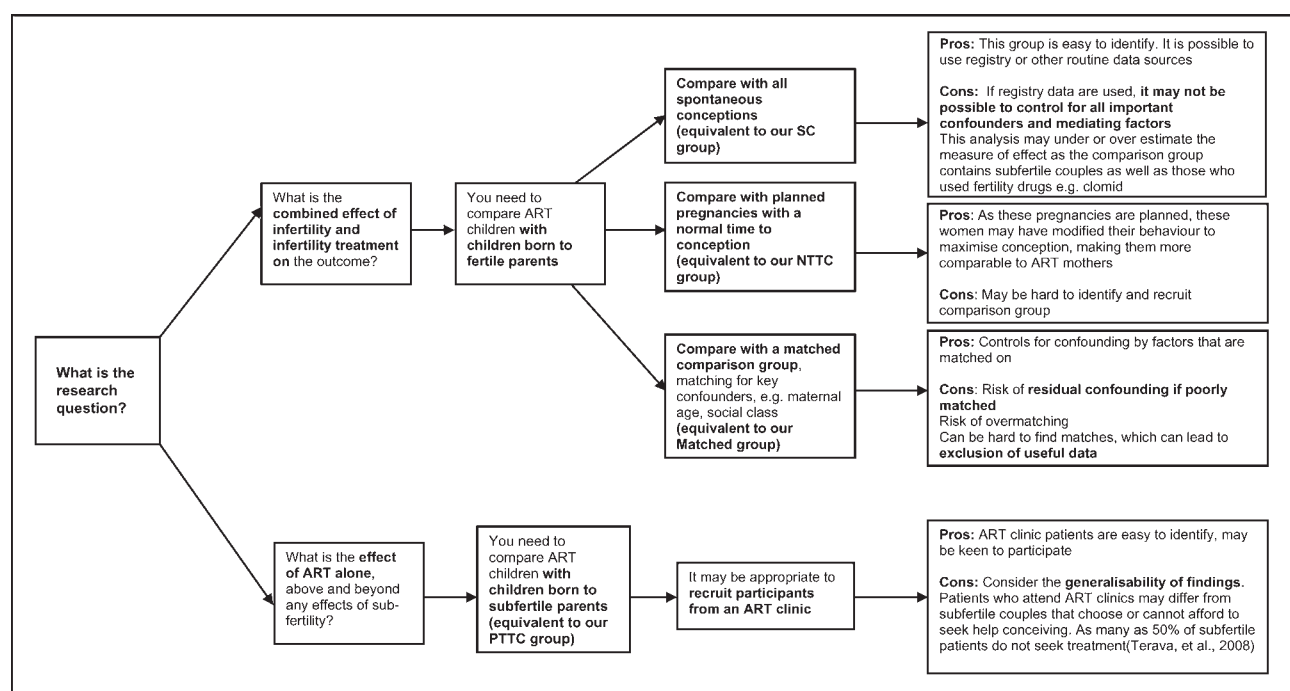


Figure 2 Summary points to consider when designing a study to examine the effect of ART on an outcome of interest.

sample of spontaneous births. However, this may under or over estimate any effects, as the comparison group also contains some subfertile couples. Studies using registry data to compare ART infants to all spontaneous conceptions may also be unable to control for all confounding factors, because the pertinent data were not collected. This means that the measure of effect may be subject to residual confounding. Our example demonstrates the mean BAS-NV score is 5.67 (2.61, 8.72) points higher in ART children compared with the general population group, and is statistically significant ($P < 0.05$). When adjusted for commonly available confounding factors this effect, although diminished, remains statistically significant. Only after controlling for additional confounding factors that are not usually available in routine data, did the advantageous effect of ART disappear. Using registry data alone would generally not provide sufficient information to be able to control for these effects and so the analyses may result in a potentially spurious finding.

A second possible comparison group is children born to women who planned their pregnancy and who conceived within a year; our ART/NTTC analysis uses such a 'fertile' group. This analysis compares the outcome in infants born after infertility and ART to those who were born to parents experiencing neither condition. Couples who are trying to conceive may modify their behaviour to improve their chances of conception; this is especially true of ART patients. By comparing ART infants to infants from planned pregnancies, we reduce any such differences. Although this comparison group provides some methodological benefits, there are also disadvantages. Infants may be harder to identify and recruit, and if the analysis is based on an existing cohort study, as ours was, a high proportion of the population will be excluded from these analyses, thus reducing statistical power.

A final approach is to conduct a matched cohort study; matching the ART and non-ART groups on key confounding factors. This method effectively removes the risk of confounding by the matched variables. Our ART/MC analysis matched baby's sex, mother's age and socioeconomic status. Consequently we cannot present a true unadjusted effect since the technique automatically controls for the matching variables. This explains why the unadjusted result for the ART/MC analysis is different from the unadjusted results for the other analyses. This method is a very efficient way of controlling for confounding. However, the disadvantages may outweigh the advantages of this design. Poor matching may result in residual confounding, whereas overmatching can inadvertently match on a factor that you are interested in, and the effects of the matched factors cannot be examined.

A design that matches on many factors can lead to the exclusion of valuable data on exposed infants, since it may be impossible to find enough suitable matched infants in the comparison group. A study by Koudstaal *et al.* (2000) illustrates how matching can be limiting: IVF twins were matched with spontaneously conceived twins on seven different factors, but one third of IVF twins had to be excluded since no matches could be found. Matched studies are labour intensive and thus generally small (Sutcliffe *et al.*, 1995; Koudstaal *et al.*, 2000) potentially resulting in a lack of statistical power. We matched on a 2:1 ratio since we were unable to match more than two comparison children for all ART infants and did not want to exclude any infants. A sample size calculation suggests that our ART/MC analysis has only 40% power to detect a true difference in mean BAS-NV of the two groups, and therefore is underpowered. For many rare outcomes

even large studies may have inadequate sample size: for example, it has been estimated that to detect a doubling in the risk of a birth defect that has 1% prevalence in the spontaneous conception group with 80% power at the 5% level of significance that 1491 ART infants and 7455 spontaneously conceived children would be needed (Kurinczuk *et al.*, 2004).

In practice, there appear to be two sensible options. If the objective is to explore the combined effect of infertility and ART on an outcome, then the simplest approach would be to compare to all other spontaneous births; the results then allow statements of comparison with all births in general. However, if the data are available, a comparison group of planned, normal time to conception infants would be optimal as this would allow comparison of like with like, and removes the complicating issue of unplanned and unwanted pregnancies. Whereas, if the aim is to examine the effect of ART over and above the effect of infertility, then a subfertile comparison group is needed.

Strengths and limitations

The MCS is a nationally representative cohort that included questions on pregnancy planning, time to conception, fertility treatment and developmental outcomes. The large dataset allows exploration of the effect of using different comparison groups. Access to detailed information on confounding factors allowed us to examine the effects of commonly available covariates, and look at the effects of confounders and mediating factors for which data are rarely available.

All data on conception, pregnancy and early life experiences were self-reported, and therefore may be prone to poor recall. However, it is unlikely that women would fail to recall ART or difficulty conceiving, so our exposed and unexposed groups are likely to be robust. There was a good response rate for the second stage of data collection (80% of families), though missing data on BAS-NV or key confounders led to the exclusion of 3793/14 376 singletons (26%).

The BAS-NV score is a well recognized, validated instrument designed specifically for the British population, and is intended to assess development in this age group. Though one of a battery of tests, it was also designed to stand alone. BAS-NV assesses only one aspect of cognitive development, namely expressive language abilities, and it should be noted that ART may potentially have a different effect on other aspects of cognitive development. The tests were conducted in the child's home, and though not purposely 'blinded' to the ART status of the infants, it is unlikely that the testers were aware of the method of conception for the children. Our analysis is one of few that looks at neurodevelopmental effects in children past toddler years: previous studies have examined younger children (4 months to 2.5 years), while this cohort was on average 3.1 years old at testing.

Conclusions

ART is associated with an increase in cognitive developmental test scores at 3 years of age, but this apparent advantage can be explained by the characteristics of the women treated with ART. Further work exploring the effects of ART on neurodevelopment in older children is needed. We have highlighted some key methodological considerations for researchers who are planning to study the effect of infertility and

ART on childhood outcomes. The choice of comparison group is important and should be determined by the research question that one wishes to address. Different data sources have important strengths and weaknesses which should be considered at the design stage; in particular the absence of data on confounding and mediating factors may lead to spurious conclusions about the nature of the causal relationship even in the largest studies. Careful consideration of whether a variable should be treated as a potential confounder or mediating factor can help to illuminate the underlying mechanism driving an observed association between an exposure and an outcome.

Acknowledgements

We would like to thank the Millennium Cohort Study families for their time and cooperation, as well as the Millennium Cohort Study team at the Institute of Education.

Funding

This project is supported by a research grant from the Medical Research Council.

References

- Agarwal P, Loh SK, Lim SB, Sriram B, Daniel ML, Yeo SH, Heng D. Two-year neurodevelopmental outcome in children conceived by intracytoplasmic sperm injection: prospective cohort study. *BJOG* 2005;**112**:1376–1383.
- Barnes J, Sutcliffe AG, Kristoffersen I, Loft A, Wennerholm U, Tarlatzis BC, Kantaris X, Nekkebroeck J, Hagberg BS, Madsen SV et al. The influence of assisted reproduction on family functioning and children's socio-emotional development: results from a European study. *Hum Reprod* 2004;**19**:1480–1487.
- Bowen JR, Gibson FL, Leslie GI, Saunders DM. Medical and developmental outcome at 1 year for children conceived by intracytoplasmic sperm injection. *Lancet* 1998;**351**:1529–1534.
- British Medical Association. In: Carter D (ed). *Smoking and Reproductive Life: The Impact of Smoking on Sexual, Reproductive and Child Health*. London: British Medical Association, 2004.
- Buck Louis GM, Schisterman EF, Dukic VM, Schieve LA. Research hurdles complicating the analysis of infertility treatment and child health. *Hum Reprod* 2005;**20**:12–18.
- Ceelen M, van Weissenbruch MM, Vermeiden JP, van Leeuwen FE, Delemarre-van de Waal HA. Growth and development of children born after in vitro fertilization. *Fertil Steril* 2008;**90**:1662–1673.
- Dex S, Joshi H. *Children of the 21st Century, The UK Millennium Cohort Study Series*. Bristol: The Policy Press, University of Bristol, 2005.
- Doyle P, Bunch KJ, Beral V, Draper GJ. Cancer incidence in children conceived with assisted reproduction technology. *Lancet* 1998;**352**:452–453.
- Elliott CD, Smith P, McCulloch K. *British Ability Scales Second Edition (BAS II): Technical Manual*, 2nd edn. London: NferNelson, 1997.
- Ericson A, Nygren KG, Olausson PO, Kallen B. Hospital care utilization of infants born after IVF. *Hum Reprod* 2002;**17**:929–932.
- Golombok S, Cook R, Bish A, Murray C. Families created by the new reproductive technologies: quality of parenting and social and emotional development of the children. *Child Dev* 1995;**66**:285–298.
- Hansen K. *Millennium Cohort Study First, Second and Third Surveys: A Guide to the Datasets*. London: Centre for Longitudinal Studies, Institute of Education, University of London, 2008.
- Hansen M, Kurinczuk JJ, Bower C, Webb S. The risk of major birth defects after intracytoplasmic sperm injection and in vitro fertilization. *N Engl J Med* 2002;**346**:725–730.
- Hansen M, Colvin L, Petterson B, Kurinczuk JJ, de Klerk N, Bower C. Admission to hospital of singleton children born following assisted reproductive technology (ART). *Hum Reprod* 2008;**23**:1297–1305.
- Helmerhorst FM, Perquin DA, Donker D, Keirse MJ. Perinatal outcome of singletons and twins after assisted conception: a systematic review of controlled studies. *Brit Med J* 2004;**328**:261.
- Hvidtjorn D, Schieve L, Schendel D, Jacobsson B, Svaerke C, Thorsen P. Cerebral Palsy, Autism Spectrum Disorders, and Developmental Delay in Children Born After Assisted Conception: A Systematic Review and Meta-analysis. *Arch Pediatr Adolesc Med* 2009;**163**:72–83.
- Kallen B, Finnstrom O, Nygren KG, Olausson PO. In vitro fertilization in Sweden: child morbidity including cancer risk. *Fertil Steril* 2005;**84**:605–610.
- Klemetti R, Sevon T, Gissler M, Hemminki E. Health of children born as a result of in vitro fertilization. *Pediatrics* 2006;**118**:1819–1827.
- Klip H, Burger CW, de Kraker J, van Leeuwen FE. Risk of cancer in the offspring of women who underwent ovarian stimulation for IVF. *Hum Reprod* 2001;**16**:2451–2458.
- Knoester M, Helmerhorst FM, Vandenbroucke JP, van der Westerlaken LA, Walther FJ, Veen S. Cognitive development of singletons born after intracytoplasmic sperm injection compared with in vitro fertilization and natural conception. *Fertil Steril* 2008;**90**:289–296.
- Koivurova S, Hartikainen AL, Sovio U, Gissler M, Hemminki E, Jarvelin MR. Growth, psychomotor development and morbidity up to 3 years of age in children born after IVF. *Hum Reprod* 2003;**18**:2328–2336.
- Koudstaal J, Bruinse HW, Helmerhorst FM, Vermeiden JP, Willemsen WN, Visser GH. Obstetric outcome of twin pregnancies after in-vitro fertilization: a matched control study in four Dutch university hospitals. *Hum Reprod* 2000;**15**:935–940.
- Kurinczuk JJ, Hansen M, Bower C. The risk of birth defects in children born after assisted reproductive technologies. *Curr Opin Obstet Gynecol* 2004;**16**:201–209.
- Leunens L, Celestin-Westreich S, Bonduelle M, Liebaers I, Ponjaert-Kristoffersen I. Cognitive and motor development of 8-year-old children born after ICSI compared to spontaneously conceived children. *Hum Reprod* 2006;**21**:2922–2929.
- Lidegaard O, Pinborg A, Andersen AN. Imprinting diseases and IVF: Danish National IVF cohort study. *Hum Reprod* 2005;**20**:950–954.
- Middelburg KJ, Heineman MJ, Bos AF, Hadders-Algra M. Neuromotor, cognitive, language and behavioural outcome in children born following IVF or ICSI-a systematic review. *Hum Reprod Update* 2008;**14**:219–231.
- Oakley L, Doyle P, Maconochie N. Lifetime prevalence of infertility and infertility treatment in the UK: results from a population-based survey of reproduction. *Hum Reprod* 2008;**23**:447–450.
- Pianta RC. *Child-Parent Relationship Scale (Unpublished measure)*. Charlottesville, VA: University of Virginia, 1995.
- Pinborg A, Loft A, Schmidt L, Andersen AN. Morbidity in a Danish national cohort of 472 IVF/ICSI twins, 1132 non-IVF/ICSI twins and 634 IVF/ICSI singletons: health-related and social implications for the children and their families. *Hum Reprod* 2003;**18**:1234–1243.
- Place I, Englert Y. A prospective longitudinal study of the physical, psychomotor, and intellectual development of singleton children up to 5 years who were conceived by intracytoplasmic sperm injection

- compared with children conceived spontaneously and by in vitro fertilization. *Fertil Steril* 2003;**80**:1388–1397.
- Plewis I. *Millennium Cohort Study First Survey: technical report on sampling*. London: Centre for Longitudinal Studies, 2007a.
- Plewis I. Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study. *Int J Soc Res Methodol* 2007b;**10**: 325–334.
- Saunders K, Spensley J, Munro J, Halasz G. Growth and physical outcome of children conceived by in vitro fertilization. *Pediatrics* 1996;**97**: 688–692.
- StataCorp. *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP, 2007.
- Stromberg B, Dahlquist G, Ericson A, Finnstrom O, Koster M, Stjernqvist K. Neurological sequelae in children born after in-vitro fertilisation: a population-based study. *Lancet* 2002;**359**:461–465.
- Sutcliffe AG, D'Souza SW, Cadman J, Richards B, McKinlay IA, Lieberman B. Minor congenital anomalies, major congenital malformations and development in children conceived from cryopreserved embryos. *Hum Reprod* 1995;**10**:3332–3337.
- Sutcliffe AG, Saunders K, McLachlan R, Taylor B, Edwards P, Grudzinskas G, Lieberman B, Thornton S. A retrospective case-control study of developmental and other outcomes in a cohort of Australian children conceived by intracytoplasmic sperm injection compared with a similar group in the United Kingdom. *Fertil Steril* 2003;**79**:512–516.
- Terava AN, Gissler M, Hemminki E, Luoto R. Infertility and the use of infertility treatments in Finland: prevalence and socio-demographic determinants 1992–2004. *Eur J Obstet Gynecol Reprod Biol* 2008; **136**:61–66.
- Wennerholm UB, Albertsson-Wikland K, Bergh C, Hamberger L, Niklasson A, Nilsson L, Thiringer K, Wennergren M, Wikland M, Borres MP. Postnatal growth and health in children born after cryopreservation as embryos. *Lancet* 1998;**351**:1085–1090.

Submitted on May 22, 2009; resubmitted on June 29, 2009; accepted on July 2, 2009