Research hurdles complicating the analysis of infertility treatment and child health

G.M. Buck Louis1,4, E.F. Schisterman1, V.M. Dukic2 and L.A. Schieve3

1Epidemiology Branch, Division of Epidemiology, Statistics & Prevention Research, National Institute of Child Health & Human Development, The National Institutes of Health, Department of Health & Human Services, 6100 Executive Boulevard, Room 7B03, Rockville, MD 20852, 2Department of Health Studies, Room W-260, University of Chicago, Chicago, IL, and 3Division of Reproductive Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA, USA

4To whom correspondence should be addressed. E-mail: gb156i@nih.gov

Research aimed at the empirical evaluation of infertility treatment including assisted reproductive technologies (ART) on child health and development is hampered by investigators’ inability to methodologically separate possible treatment effects from underlying fecundity impairments. While the literature continues to identify ART as a risk factor for many child health outcomes, less attention has been paid to the methodologic rigor needed to answer this question. We identify aspects of fecundity and the nuances of medical practice that need to be considered and captured when designing epidemiologic investigations aimed at assessing ART and child health. These include: (i) the use of prospective study designs in which the unit of analysis (cycle versus individual versus couple) is defined; (ii) data collection on relevant time-varying covariates at, before and during treatment; and (iii) the use of statistical techniques appropriate for hierarchical data and correlated exposures. While none of these issues in and by itself is unique to ART research, attention to these issues has been lacking in much of the published research limiting our ability to evaluate health consequences for children. Longitudinal studies of children conceived with ART will benefit from attention to these issues and, hopefully, produce answers to lingering questions about safety.

Key words: assisted reproductive technologies/child health/correlated outcomes/design/hierarchical models

Introduction

There is an accumulating body of research suggesting that infertility treatment may adversely affect the health and development of children so conceived. Singleton infants born to couples undergoing assisted reproductive technologies (ART) have been reported to be at increased risk for decrements in gestation and birth size (Australian In vitro Fertilization Collaborative Group, 1985; Rufat et al., 1994; Koudstaal et al., 2000; Schieve et al., 2002; Schieve et al., 2004a), select birth defects (Bergh et al., 1999; Wennerholm et al., 2000; Ericson and Kallen, 2001; Hansen et al., 2002) and developmental disabilities (Stromberg et al., 2002), although equivocal findings exist (Anthony et al., 2002; Van Steirteghem et al., 2002). Reviews have summarized the health outcomes of concern for children (Helmerhorst et al., 2004; Schieve et al., 2004b). A recent random effects meta-analysis comparing 12,283 singleton infants conceived with IVF with 1.9 million spontaneously conceived infants reported treatment to confer a significantly two-fold increased risk of perinatal mortality, preterm delivery, low birth weight and small-for-gestational age (Jackson et al., 2004) after controlling for maternal age and parity.

Epidemiologists and biostatisticians have an opportunity to lend expertise in designing study protocols responsive to the many questions about ART effects. Technically speaking, ART treatment includes procedures that involve the handling of human sperm and eggs outside the body for the expressed purpose of creating a pregnancy (i.e. transcervical embryo transfer, ICSI, gamete and zygote intrafallopian transfer, frozen embryo transfer, and donor embryo transfer). Timely answers are needed as the number of treated couples steadily increases along with the number of live births. For example, ART treatment procedures have increased approximately 20-fold in the United States since 1986, accompanied by a 100-fold increase in the number of ART conceived infants (Medical Research International and The American Fertility Society Special Interest Group, 1988; CDC et al., 2003). This percentage is likely to increase as treatment efficacy continues to improve and, in the United States, as health-care coverage becomes increasingly mandated (RESOLVE and The National Infertility Association, 2003).

We have identified key methodological nuances that challenge investigation regarding ART and child health while offering strategies for addressing them. This is done in a question–answer format beginning with study design through...
Can we study the effect(s) of ART treatment on child health?

Yes. The relatively short interval between initiation of ART treatment and time to event (i.e. implantation, pregnancy, live birth) underscores the utility of prospective design while ensuring a temporal relation between exposure and outcome. Some retrospective observational designs might be more efficient for rare outcomes such as Beckwith–Wiedemann syndrome or in offering an interim assessment of long-term outcomes such as the future fertility of children conceived with ART treatment. These designs are less desirable, overall, because of the potential for bias, imprecision with regard to temporal ordering and correlated exposures. Both descriptive (e.g. case series and cross-sectional) and retrospective (i.e. case control) study designs have been successfully used in previous research initiatives aimed at assessing the impact of ART on infant health. The complexity of this question and underlying methodological nuances underscore the role for prospective study as outlined in this paper. In particular, two key issues are: (i) deciding on the right unit of analysis; and (ii) defining what time scale should be used.

With respect to the first consideration, investigators will need to decide a priori whether the research question is aimed exclusively at female, male or couple-based infertility and accompanying treatment modalities. This is an important decision that impacts study design and the interpretation of results. Investigators need to a priori determine whether the unit pertains to a cycle, individual (woman or man) or couple, in a manner consistent with the study goals and research objectives. This can be a challenging decision, as seen with ICSI and infant outcomes. Although ICSI was originally developed to overcome fertilization failure in couples with male factor infertility, ICSI is increasingly used to treat couples irrespective of male factor.

In 1995, ICSI was used in 11% of ART procedures using embryos from the freshly fertilized patient’s eggs that were reported to the United States ART registry increasing to 49% in 2001 (CDC et al., 1997). Thus, even for a very select ART procedure such as ICSI that originated as a treatment for male factor, there is now much more diversity with respect to fecundity among couples using ICSI. This example underscores the heterogeneity among infertile couples utilizing ART treatments in contemporary clinical practice. If there is any reason to suspect parental fecundity or treatment effects, a couple-based approach is needed. Given the dearth of empirical evidence on the determinants of male and female fecundity (Buck et al., 1997), every effort should be expended to first consider a couple-based cohort design to ensure effects are not inadvertently designed away.

With respect to the time consideration, investigators can define time (and hence time-varying covariates) as calendar time (e.g. age, duration of infertility) or in terms of treatment cycles. Time needs to be specified upon designing a study according to the study’s hypotheses. For example, timing of maternal and/or paternal exposures is essential for determining imprinting outcomes such as genetic and chromosomal constitution of the sperm sample used for ICSI. The need to identify conception as a part of ART treatment enables researchers to consider and measure day-specific exposures. Considerable care should be taken not to interchange time or the unit of analysis (e.g. flip from calendar to cycle time or flip from couples to women) in the data analysis stage without an appropriate a priori analytic plan. For example, changing the time scale from cycle to calendar time may result in loss of power because some couples may have fewer cycles in a given time period than others. Changing from calendar to cycle time may result in emergence of time/age-related confounders that could bias the analysis, if not properly accounted for.

Can we differentiate ART treatment effects from underlying fecundity impairments?

No. The ideal study design for answering this question is a randomized clinical trial, which is not possible given that it would require administering ART to fecund couples. However, it is possible to evaluate ART effects among infecund couples receiving treatment. To better understand our inability to separate infertility from an ART treatment effect, we offer the framework of a randomized trial as illustrated in Figure 1.

![Figure 1. Conceptual framework for designing studies focusing on the effects of ART treatment on child health and development.](https://academic.oup.com/humrep/article-abstract/20/1/12/671494)
As shown in Figure 1, $p_1$ denotes the proportion of study outcomes among infecund couples who achieved pregnancy with ART. Similarly, $p_2$ refers to the proportion of study outcomes among infecund couples who achieved pregnancy without ART treatment, while $p_1$ and $p_4$ represent the proportions of study outcomes among fecund couples who achieve pregnancy with and without ART, respectively. We can conceptualize a framework for empirically estimating effects, illustrated as relative risks (RR), with respect to three questions: (i) the treatment effect on human development among infecund couples (RR$_1 = p_1/p_2$); (ii) the infecundity effect on human development for couples undergoing treatment (RR$_2 = p_1/p_3$); and (iii) the joint effect of infecundity and related treatment on human development (RR$_3 = p_1/p_4$). For observational or non-randomized studies, we will need a relatively large cohort of couples trying to become pregnant to attempt answering these questions.

Absence of population estimates for $p_2$ makes it difficult to estimate RR$_1$. The probability of becoming pregnant without medical intervention among infecund couples (those trying for >12 months without success) is virtually unknown as we are unaware of any population-based incidence figures on resolved infertility. Moreover, couples that do eventually conceive without assistance are probably different with respect to fecundity and exposure profiles from those who do not. However, clinical findings of couples that present to medical centres for diagnostic testing after failure to conceive suggest that approximately 14% of untreated infertile couples will eventually give birth (Collins et al., 1995).

Another approach with observational designs is to estimate the effect of infecundity alone among couples undergoing treatment (RR$_3$). Although limited anecdotal evidence suggests that ART may be used by women or couples with no indication of infecundity (e.g. single women who elect to use ART after no or only a short time trying to conceive naturally or with insemination), there is currently no way to identify such cases in many registries, including the United States ART registry. While there may be very limited circumstances under which this research question might be addressed, it is virtually impossible to do so at the population level. Moreover, it is likely that such cases make a small contribution to the total population of couples currently receiving ART treatment. Thus, we are realistically left with estimating RR$_3$ in an observational setting.

As stated before, RR$_3$ is the compound effect of infecundity and treatment. From a public health perspective, this is the entity of utmost concern, because the compound effect has direct relevance for couples trying to weigh the total risks associated with conception after ART, whether these risks stem from the treatment or the underlying fecundity impairment. We do, however, recognize that the compound effect is limited in informing aetiology.

**How can we evaluate the effects of ART treatment on child health?**

Two strategies for prospective recruitment of couples that will eventually undergo ART can be used to answer this question. The first is the recruitment of couples from the population or a target subgroup prior to attempting pregnancy or after failing to become pregnant during a specified period of time, such as 6 months. This approach is supported by empirical evidence that suggests approximately 65–85% of couples attempting pregnancy will do so within 6 months (Bonde et al., 1998; Wang et al., 2003). Moreover, investigators can collect baseline information and biospecimens prior to entering treatment along with longitudinal information on treatment and relevant covariates once treatment is initiated. This approach would allow investigators to capture information on the full range of treatments received prior to ART, albeit a number of important methodologic challenges and limitations. While past authors have demonstrated the validity and reliability of time-to-pregnancy (TTP) information recalled by men and women for 20+ years (Baird et al., 1991; Zielhuis et al., 1992; Joffe et al., 1993; 1995; Olsen et al., 1998), other authors have reported biases associated with retrospectively collected TTP such as a woman’s behaviours, intentions about becoming pregnant or ability to recognize pregnancy (Weinberg et al., 1994).

Other limitations associated with the use of couples failing to become pregnant is the relatively large initial sample size required to enroll a sufficient number of ART births, especially when the study objective is to evaluate rare infant outcomes such as birth defects. An insufficient sample size with respect to statistical power may lead to an erroneous conclusion about the absence of an ART treatment effect (type II error).

Another prospective cohort approach, albeit less desirable, is to recruit couples upon first seeking services for perceived infecundity. Practical challenges include the receipt of a range of ‘infertility’ services from various health providers. Some women may be placed on ovulation induction agents prior to a full fecundity evaluation. These issues underscore the potential for information or misclassification bias along with other study limitations such as marked heterogeneity with respect to treatment and the need for retrospective collection of covariate data up to the time medical care was sought. This is particularly troublesome if exposure patterns change either temporally (e.g. seasonal use of pesticides or changes in cigarette smoking when perceiving conception difficulties), or purposefully if couples alter behaviours after experiencing failure (e.g. taking dietary supplements or vitamins).

**What are possible sampling strategies for evaluating the effects of ART treatment on child health?**

In the United States there are few sampling frameworks for ascertaining populations at risk for fecundity or fertility impairments, necessitating the need for creativity on the part of investigators. This has led to the use of convenient or clinically based samples for assessing ART treatment effects. The absence of population-based samples limits external validity or generalizability of study results. As such, the study results cannot be used in empirically based clinical decision-making beyond the treated individuals comprising the sample.
Choice of sampling strategy is impacted by the observation that a large percentage of infecund couples do not seek medical care. The prevalence of medical care-seeking behaviour for primary infecundity ranges from 32% to 92%, and from 22% to 79% for secondary infertility (Schmidt et al., 1995). Approximately 44% of US women reporting fecundity impairments stated that they had not sought medical services (Stephen and Chandra, 1998). Care-seeking behaviour has been reported to be associated with woman’s age, education, income, duration of marriage and nulliparity (Rachootin and Olsen, 1981; Hirsh and Mosher, 1987), raising concern about potential selection bias. Even less is known about care-seeking behaviours of men, despite concerns about declining male fecundity or the so-called testicular dysgenesis hypothesis (Skakkebaek et al., 2001). Perhaps, differences in care-seeking behaviour will be minimized as health insurance providers routinely include ART services.

Implicit with sample selection is the need for sufficient sample size to minimize erroneous interpretation of study results (type II error). The unit of analysis (i.e. treatment cycle versus woman versus couple) is critical not only for determining the sampling framework, but also to ensure the analysis is appropriately statistically powered. Again, less prevalent child health outcomes such as birth defects or cancer must be designed with a sufficient sample size.

Which health end points are appropriate for study? Can more than one end point be evaluated in a single study?

A spectrum of study end points can be studied in most prospective cohort designs. While many outcomes are well defined and suitable for measurement (e.g. number of oocytes retrieved or fertilized, number of live births), others require careful operational definitions and are subject to measurement error. For example, pregnancy can be defined as a biochemical or HCG detected pregnancy, or as a clinical pregnancy as evidenced by the presence of a gestational sac or heartbeat. Pregnancies detected by HCG need to exclude false-positive detections associated with a residual treatment effect. Clinical pregnancy as an end point automatically excludes approximately two-thirds of all post-implantation HCG pregnancies and may introduce the potential for biased findings (Wilcox et al., 1988; Wang et al., 2003). The interplay between early pregnancy loss and fecundability in the following cycle has received limited study. Previous authors have found that early (HCG detected) pregnancy loss is associated with a significant increase in the odds of total conception, clinical pregnancy and subsequent early loss (Wilcox et al., 1988; Wang et al., 2003). These findings further support the role of prospective studies inclusive of a spectrum of end points that may be suitable for the eventual delineation of the effect of prior pregnancy outcome on subsequent outcome regardless of mode of conception. Thus, varying definitions for outcomes will impact the study findings and their interpretation. The study goals should a priori dictate which end points to study and the methodology should delineate operational definitions for all outcomes.

Relevant developmental end points include: plurality of birth, fetal growth and development, birth size, birth defects, developmental disabilities, and cancer. Other aspects of human development might include sexual maturation and puberty, necessitating the need for following children through adolescence. Prospective design also enables investigators to capture incident cases or affected infants in lieu of prevalent cases that might be ascertained in a retrospective study design.

Added challenges in assessing developmental end points pertain to the issue of multiple births. In the United States, an estimated 37% of ART pregnancies that progress far enough to detect fetal hearts are classified as multiple gestations and 35% of live born deliveries are multiples. Considering total infants born from ART, more than half are twins or higher order, 18 times the proportion reported for the general United States population (Wright et al., 2003). Given the inherent perinatal sequelae associated with multiples (ESHRE Capri Workshop Group, 2000), the need to clearly identify multiple pregnancies and not just multiple births is evident. This can be a difficult task, as many early multiple gestations are reduced to a singleton clinical pregnancy (Salihu et al., 2003). Recent data from the United States registry of ART procedures indicate that approximately 9% of singleton live births conceived from ART were from pregnancies with two or more fetal hearts observed on early ultrasound (Schieve et al., 2004a). The fetal death of a co-twin is a known risk factor for neonatal complications and morbidity in the remaining twin (Skrablin et al., 1994), especially in the presence of a twin-to-twin transfusion (Saito et al., 1999).

How and when should ART treatment data be captured for analysis?

Timing of data collection is inherently a function of the unit of analysis (cycle versus individual versus couple) and measure of time (calendar time versus age versus treatment duration), not to mention the plethora of relevant clinical and behavioural factors known to impact on the probability of conception and birth. The longstanding recognition of critical windows important for human development further underscores the importance of specifying time-varying covariates (Wilson, 1965; Selevan et al., 2000), including periconception (Morford et al., 2004). For ART couples alone, investigators have a unique ability to identify conception and, hence, gestational age-specific exposures. Research efforts currently aimed at determining day-specific probabilities of conception among couples attempting pregnancy offer further promise for the eventual analysis of time-specific exposures (Ferreira-Poblete, 1997; Roysen and Ferreira, 1999). One of the biggest challenges facing research in this area is to capture the treatment data including individual clinic nuances when administering treatment protocols to couples (e.g. adding aspirin use to treatment protocols). The added challenge will be to capture the specific timing of the treatment in the context to other exposures the couple may face. In essence, couples are exposed to a mixture of factors,
of which only one broad category is the ART treatment itself. This situation is similar to other avenues of research such as the study of environmental chemicals in the context of other physical agents (e.g. pesticides and heat) and human development.

How should data be analysed when evaluating ART treatment and child health?

Data analysis is complex on many levels, from conceptualizing the research question to devising an analytic plan appropriate to that question. Investigators are faced with interesting challenges including: (i) hierarchical nature of ART treatment; (ii) correlation patterns of exposures with respect to timing; and (iii) potential for parental interactions.

ART observations will generally not be independent, given the need to consider multiple oocytes, implantations or embryos per treatment cycle, as well as multiple treatment cycles per woman or couple. Failure to account for such dependency may lead to unpredictable bias (Dukic and Hogan, 2002; Daya, 2003). Investigators are left with either restricting treatment cycles to one per woman or to the use of analytical models capable of addressing the dependency [Bayesian methods, mixed models or generalized estimating equations (GEE)].

Without consideration of the issues of dependency and correlation, study findings may be inefficient and biased, ultimately yielding imprecise or even incorrect conclusions. It is known that clustering (form of dependency) commonly occurs in ART data, where outcomes are clustered on several levels: within menstrual or treatment cycles, within each woman and within each couple. The complex correlation structure generally observed in ART outcomes can be due to unmeasured (or omitted) level-specific covariates, such as, for example, unmeasured endometrial or embryo characteristics or omitted time-varying exposure.

Unfortunately, many investigators still report and rely on simple statistics such as t-tests or Pearson’s $\chi^2$ statistic, in spite of multiple confounders and the lack of independence of pregnancy outcomes. Such published work may mislead clinicians and, ultimately, couples undergoing care. A variety of statistical techniques exist today for analysing data in situational clustering occurs and induces dependency among the outcomes of interest. Tools for analysing dependent outcomes are well developed and implemented in a variety of standard statistical packages. For example, longitudinal data analysis techniques, GEE, (empirical) Bayesian hierarchical models and mixed models are readily available in widely used statistical software packages such as Stata or SAS. Other appropriate specialized statistical packages include MIXOR, EGRET, HLM, MLWin, or WinBUGS and BUGS (Hogan and Blazar, 2000).

Hierarchical models (also known as multi-level models, random-effects or mixed models) can account for this within-cluster correlation by assuming that each cluster has its own (unobservable) probability of a positive outcome, which is different from all other clusters, and that all cluster-specific probabilities follow some distribution. The variance of this distribution is what is known as the heterogeneity parameter, and its magnitude reflects the amount of heterogeneity among women or couples (Hogan and Blazar, 2000; Dukic and Hogan, 2002). Hierarchical models also can provide ‘cluster-specific’ estimates of treatment and other relevant covariate effects. This feature is desirable in many clinical settings when one is interested in prediction of outcomes for a given woman or a given couple.

All Bayesian models are in fact hierarchical models, where the top hierarchical level is reserved for so called ‘priors’, or summaries of knowledge elicited from previous studies or from expert opinion. This ability to formally account for prior knowledge can be extremely useful, particularly in ART settings. Considerable information on prior history is likely to exist for many couples involved in ART studies, and investigators should be able to formally include that information in the analysis.

Methods like GEE can also account for within-cluster correlation. However, they cannot produce subject-specific estimates, but instead yield population-level inference. This can be well suited for policy making, but less useful for individual clinical prediction. Other methods for handling dependent outcomes are available, such as variance-adjustment techniques (sandwich estimators) or bootstrap.

Another very important analytical issue is missing data and informative drop-out in a longitudinal study setting. Missing data mechanisms and reasons for dropout can in general be extremely complicated in an ART setting, resulting in severe selection bias for ART research. To address non-negligible dropout issues, explicit models for the drop-out mechanism or joint models for the outcome of interest and time to drop-out are needed (Hogan and Laird, 1997). Appropriate statistical input will generally be required for design and implementation of such research.

What conclusions can be reached from available and future data?

As with all epidemiological research, study findings need to be evaluated within one of the established paradigms for assessing causality (Weed, 1995). Such an approach ensures that the methods are appropriate given the design and that the results are interpreted within the study’s limitations and strengths. Two particularly worrisome observations that appear in the published literature to date with respect to the interpretation of ART treatment effects and child health are lack of statistical power for the relevant unit of analysis and limited attention to alternative explanations. With respect to power, several authors report the absence of an ART treatment effect on rare outcomes such as birth defects, despite insufficient statistical power (Sutcliffe et al., 1995; Dhont et al., 1997). The issue of limited statistical power for birth defects and developmental disabilities for many studies has recently been addressed (Buitendijk, 1999) and can be exaggerated further if a good response rate is not achieved. This concern is further amplified by the fact that ignoring dependency in ART outcomes will tend to overestimate the precision of the findings and consequently result in
overestimated statistical power. These early studies, albeit underpowered for many outcomes, were initial attempts to assess health implications of ART on children’s health. Current investigators need to build upon these efforts with the implementation of study designs appropriate for the research question and sensitive to the methodological nuances described in this paper if definitive answers are to be obtained.

As researchers, we are left answering the question about RR3 as described above: what is the compound effect of infertility and ART treatment on child health? This challenge requires the use of a standardized study protocol involving prospective data collection with various laboratory, clinical and behavioural factors, and measures of fecundity impairment. With accurate information on type and diagnostic subtype of infecundity, it will be possible to adjust statistical analyses to further assess effects.

Summary

Despite the challenges in designing research initiatives aimed at assessing the relationship between ART treatment and child health as globally defined, concerted study can be done and answers to pressing questions are possible. This will most likely require conceding that we will not be able to totally eliminate effects stemming from underlying fecundity impairments from treatment. But, we will be able to estimate child health risks among infecund couples receiving ART (or RR3, as defined in this paper as the compound effect of infecundity and ART). This alone will be a contribution to the field and offer promise for the early identification of potential reproductive and developmental toxicants at critical windows for humans.

Acknowledgements

The authors thank Dr Courtney J. Lynch for her critical review and feedback provided on the paper.

References


CDC, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology and RESOLVE (1997) 1995 Assisted Reproductive Technology Success Rates: National Summary and Fertility Clinic Reports. CDC, Atlanta, GA, USA.

CDC, American Society for Reproductive Medicine and Society for Assisted Reproductive Technology (2003) Assisted Reproductive Technology Success Rates: National Summary and Fertility Clinic Reports. CDC, Atlanta, GA, USA.


Skakkebaek NE, Rajpert-De Meyts E and Main KM (2001) Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects. Hum Reprod 15,972–978.


Submitted on May 12, 2004; accepted on September 10, 2004