Human Reproduction, Vol.34, No.3 pp. 549–557, 2019
Advanced Access publication on December 21, 2018 doi:10.1093/humrep/dey374

human reproduction

ORIGINAL ARTICLE ***Reproductive epidemiology***

Methodological approaches to analyzing IVF data with multiple cycles

Jennifer Yland[1,*], Carmen Messerlian[2], Lidia Mínguez-Alarcón[2], Jennifer B. Ford[2], Russ Hauser[1,2,3], and Paige L. Williams[1,4], for the EARTH Study Team

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA [2]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA [3]Massachusetts General Hospital Fertility Center, Department of Obstetrics and Gynecology, Harvard Medical School, Boston, MA, USA [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

*Correspondence address. Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. E-mail: JEY438@mail.harvard.edu

Submitted on August 29, 2018; resubmitted on November 21, 2018; accepted on December 14, 2018

STUDY QUESTION: Which methodological approaches are most appropriate for analyzing IVF data with multiple cycles in the context of a binary outcome?

SUMMARY ANSWER: Both mixed effect models and generalized estimating equation (GEE) modeling approaches can account for multiple IVF cycles and may reduce bias over first-cycle only approaches, but CIs were narrowest with cluster-weighted generalized estimating equation models (CWGEE).

WHAT IS KNOWN ALREADY: There is a lack of consensus among investigators regarding how to best incorporate data from multiple cycles and whether to present odds or risks in the analysis of IVF data. Failure to account for correlated outcomes within individuals and informative cluster size may lead to invalid CIs and biased estimates.

STUDY DESIGN, SIZE, DURATION: The Environment and Reproductive Health (EARTH) Study is an ongoing prospective cohort study of subfertile couples conducted at an academic medical center. This cohort was established in 2004 and follows couples seeking treatment for infertility throughout the course of their treatment and pregnancy.

PARTICIPANTS/MATERIALS, SETTING, METHODS: Women aged 18–46 years enrolled in the EARTH Study from 2004 to 2017 who initiated at least one IVF cycle were eligible. Cycle initiation was defined as beginning ovulation induction with the intent to progress through an IVF or ICSI cycle. This analysis included 442 women undergoing 642 cycles who met the study inclusion criteria. We compared the results and interpretations of log-binomial and logistic models restricting to the first cycle, as well as mixed effects models, unweighted GEE models, and CWGEE models including all cycles. This analysis was conducted for two distinct exposures: maternal age at cycle initiation, and maternal preconception urinary concentrations of di(2-ethylhexyl) phthalate (DEHP) metabolites (previously reported to be associated with a decreased probability of live birth).

MAIN RESULTS AND THE ROLE OF CHANCE: In general, the CIs were widest for mixed effects models and narrowest for CWGEE models. Further, in models evaluating the sum of urinary concentrations of DEHP metabolites ($\sum$DEHP, available for 91% of women), the point estimates were surprisingly different between the first-cycle and multiple-cycle models. We observed significant associations between maternal age and live birth in all models. However, we observed no associations between $\sum$DEHP and live birth.

LIMITATIONS, REASONS FOR CAUTION: This analysis was limited to an example dataset in which the true effect of any exposure is unknown. While this allows us to observe model performance in the context of real data, future analyses should be conducted within simulated datasets under various assumptions to further evaluate the appropriateness of each approach. In addition, we did not address differential loss to follow-up in our statistical approaches.

WIDER IMPLICATIONS OF THE FINDINGS: The use of CWGEE models should be more widely considered in the analysis of IVF data with multiple cycles per woman. The CWGEE approach is computationally simple, addresses non-ignorable (informative) cluster size, and is robust against mis-specification of the underlying covariance structure. Among the methods compared in this analysis, CWGEE models

© The Author(s) 2018. Published by Oxford University Press on behalf of the European Society of Human Reproduction and Embryology. All rights reserved.
For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

generally yielded the narrowest CIs, possibly indicating the most precise estimates. We also stress the importance of estimating risks rather than odds in the analysis of IVF data.

STUDY FUNDING/COMPETING INTEREST(S): The project was funded by Grants (R01ES022955, R01ES009718, and P30ES000002) from the National Institutes of Health. None of the authors has any conflicts of interest to declare.

Key words: IVF / ART / infertility / research methods / clustered data / informative cluster size.

Introduction

Infertility, defined as the failure to establish a clinical pregnancy after 12 months of regular, unprotected sexual intercourse, affects up to 15% of couples trying to conceive (Louis *et al.*, 2013; Thoma, *et al.*, 2013). Many couples turn to IVF for treatment. In 2015, more than 200,000 IVF cycles were performed in the USA, compared to about 60,000 in 1995 (Society for Assisted Reproductive Technology, American Society for Reproductive Medicine, 2007; ASRM, 2017). An IVF cycle begins with ovarian suppression, followed by controlled stimulation, ovulation induction, oocyte retrieval, fertilization and embryo transfer. Each cycle can fail at any step prior to or after embryo transfer.

Developing and implementing valid, appropriate statistical approaches for evaluating predictors of success in studies of IVF remains a significant challenge in ART research (Messerlian and Gaskins, 2017). However, the field lacks consensus on which approaches are most appropriate for yielding unbiased parameter estimates, especially when considering multiple cycles (Buck Louis *et al.*, 2005; Messerlian and Gaskins, 2017). Many couples receiving IVF treatment undergo multiple cycles before achieving a live birth or discontinuing treatment. In a cohort study of 6164 patients undergoing IVF in Massachusetts between 2000 and 2005, 62% had more than one IVF cycle, and the maximum number of cycles was 10 (Malizia *et al.*, 2009).

In the analysis of IVF data, it is advantageous to consider outcomes from all cycles in order to draw clinically relevant inferences and to maximize study power. However, many research publications in this area restrict analysis to the first cycle, as the inclusion of all subsequent cycles introduces greater statistical complexity as well as concern for potential bias. Cycle outcomes are likely to be correlated within each patient, and failure to account for this correlation may lead to underestimation of SEs. In addition, the number of prior cycles that a patient has undergone is likely to be informative of her probability of success. Treating all cycles equally would over-weight couples with the most severe infertility. This may lead to estimates of the per-cycle probability of success that are biased downwards, or to overestimated associations with exposure.

In this investigation, we evaluated several methods to estimate the association of exposure measures with the probability of live birth. We considered the utility of reporting risks versus odds, by comparing first-cycle log-binomial and logistic models. We further evaluated various methods to account for non-independent cycle outcomes and informative cluster size. Specifically, we considered potential sources of bias and compared the results and interpretations of mixed effects models, unweighted generalized estimating equation (GEE) models, and cluster weighted GEEs (CWGEE). We conducted these analyses within the Environment and Reproductive Health (EARTH) Study. The primary exposure was maternal age at cycle initiation, which is a well-established predictor of live birth following IVF (Malizia et al., 2009; Centers for Disease Control and Prevention *et al.*, 2017). We also compared these methods for assessing the relationship between maternal preconception urinary concentrations of di(2-ethylhexyl) phthalate (DEHP) metabolites and live birth. Phthalates are a group of plasticizers commonly found in consumer products, and exposure to these chemicals is widespread (Centers for Disease Control and Prevention, 2013). Phthalates have previously been reported to be associated with a decreased probability of live birth in the EARTH Study (Hauser *et al.*, 2016).

Materials and Methods

Study participants

The EARTH Study is an ongoing prospective cohort study conducted in collaboration with the Massachusetts General Hospital (MGH) Fertility Center and the Harvard T.H. Chan School of Public Health. This cohort was established in 2004 and follows couples seeking treatment for infertility throughout the course of their treatment and pregnancy. Data are collected on a variety of environmental, nutritional and lifestyle exposures. Details of the EARTH Study have been described previously (Messerlian *et al.*, 2018). Women who were enrolled in the EARTH cohort from 2004 to 2017 and initiated at least one IVF cycle were eligible for this study, regardless of cycle outcome or treatment discontinuation. Cycle initiation was defined as beginning ovulation induction with the intent to progress through an IVF or ICSI cycle. Women who were oocyte donors, conceived naturally or who only received IUI were excluded. We also excluded individual cycles in which patients received gamete donation or used cryo-thawed oocytes. We further excluded women ($n = 4$) with missing information on BMI, as BMI is an established predictor of IVF success (Rittenberg *et al.*, 2011). Finally, we excluded all frozen embryo transfer cycles ($n = 103$), which accounted for 14% of cycles after all other inclusion criteria were applied. Note that in the latter cases, individual cycles were excluded, but women may have had other initiated cycles included in the analysis. At the time of recruitment, trained research staff explained all procedures and answered relevant questions. The study was approved by the Human Studies Institutional Review Boards of MGH, the Harvard T.H. Chan School of Public Health, and the Centers for Disease Control and Prevention (CDC). Participants signed an informed consent after the study procedures were explained by a research nurse and all questions were answered.

Clinical data and covariates

Clinical data regarding each IVF cycle were abstracted from each patient's medical records by trained research staff. These data included treatment cycle number, treatment protocol (luteal phase agonist, flare or antagonist), fertilization protocol (IVF or ICSI) and pregnancy outcome. In addition, each study participant was assigned an infertility diagnosis by their treating physician, according to the Society for Assisted Reproductive Technology

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

Table I Demographic and clinical characteristics among women in the Environment and Reproductive Health Study enrolled between 2004 and 2017.

Woman-specific characteristics	Maternal age *n* (%) (*N = 442 women*)	∑DEHP *n* (%) (*N = 401 women*)
Age at study entry (years: Mean ± SD)		
	35 ± 4	35 ± 4
BMI at study entry (kg/m^2 Mean ± SD)		
	24 ± 4	24 ± 4
Smoking		
Current smoker	11 (2)	9 (2)
Past smoker	109 (25)	97 (24)
Never smoker	322 (73)	295 (74)
Race		
Caucasian	367 (83)	335 (84)
Black/African American	15 (3)	13 (3)
Asian	42 (10)	37 (9)
Other	18 (4)	16 (4)
Primary SART Diagnosis at study entry		
Female factor	137 (31)	124 (31)
Male factor	137 (31)	128 (32)
Unexplained	168 (38)	149 (37)
Infertility at study entry		
Primary	270 (61)	246 (61)
Secondary	170 (38)	153 (38)
Unsure	2 (1)	2 (1)
Number of IVF cycles		
Mean ± SD	1.5 ± 0.7	1.5 ± 0.8
Range	1–5	1–5
Cycle-specific characteristics	*(N = 642 cycles)*	*(N = 575 cycles)*
Treatment Protocol		
Luteal phase agonist	417 (65)	379 (66)
Flare	124 (19)	107 (19)
Antagonist	101 (16)	89 (15)
Fertilization protocol	*(n = 602)*	*(n = 543)*
ICSI	336 (56)	309 (57)
Traditional IVF	266 (44)	234 (43)
Cycle outcome		
No oocytes retrieved	40 (6.2)	32 (5.6)
Fertilization failure	16 (2.5)	13 (2.3)
Embryos not transferred	24 (3.7)	18 (3.1)
Implantation failure[a]	229 (35.7)	210 (36.5)
Chemical pregnancy[b]	40 (6.2)	36 (6.3)
Ectopic pregnancy	6 (0.9)	5 (0.9)
Spontaneous abortion	46 (7.2)	40 (7.0)
Therapeutic abortion	4 (0.6)	4 (0.7)
Stillbirth	3 (0.5)	3 (0.5)
Live birth	234 (36.5)	214 (37.2)

DEHP, di(2-ethylhexyl) phthalate; SART, Society for Assisted Reproductive Technology.
[a]Implantation failure was defined as a negative pregnancy test (bhCG < 6 mIU/ml) 17 days following embryo transfer or insemination.
[b]Chemical pregnancy was defined as implantation with no subsequent clinical pregnancy.

(SART) (Centers for Disease Control and Prevention *et al.*, 2013). Pertinent demographic data and prior pregnancy were obtained via a baseline questionnaire at study entry. Upon enrollment, a member of the research study staff measured each patient's height and weight. BMI was calculated as weight (kg) per height squared (m^2).

Urinary phthalate metabolite concentrations

In a subset of EARTH participants with available measures, we also considered associations of live birth with the sum of urinary concentrations of DEHP metabolites, or ∑DEHP, a potentially modifiable exposure. Urine spot samples were collected at study entry, as well as twice during each IVF cycle: between Days 3 and 9 of the gonadotrophin phase and on the day of oocyte retrieval, as described previously (Silva *et al.*, 2007; Hauser et al., 2016). The DEHP metabolites included mono(2-ethylhexyl) phthalate (MEHP), mono(2-ethyl-5-hydroxyhexyl) phthalate (MEHHP), mono(2-ethyl-5-oxohexyl) phthalate (MEOHP) and mono(2-ethyl-5-carboxypentyl) phthalate (MECPP). Concentrations below the limit of detection (LOD) were replaced with the LOD divided by $\sqrt{2}$ (Hornung and Reed, 1990; Hauser et al., 2016). We adjusted the metabolite concentrations by specific gravity and calculated cycle-specific metabolite concentrations by taking the geometric mean of the two samples from each IVF cycle (Hauser et al., 2016). The molar sum of DEHP metabolites (∑DEHP) was calculated by dividing each metabolite concentration by its molecular weight and then summing across metabolites (Hauser et al., 2016).

Statistical analysis

We compared several statistical methods for evaluating the association of exposures with live birth, both restricting to the first IVF cycle and including all cycles. The two exposures of interest were quartiles of maternal age at cycle initiation, available for all women, and ∑DEHP, available for 91% of women. Each of these exposures was evaluated for each statistical method. Patient-specific demographics, as well as cycle-specific characteristics, were examined and reported in the full cohort and in the ∑DEHP subsample.

We considered two approaches for evaluating the association between exposure and live birth, restricted to the first cycle. First, relative risks were obtained with a log-binomial model; second, odds ratios (ORs) were estimated with logistic regression. The log-binomial model provides a more relevant and intuitive estimate compared to the logistic model for prospective cohorts with outcomes that are not rare. Specifically, the OR is not a readily interpretable measure, and will be farther from the null compared to the risk ratio (RR) when the outcome is common (e.g. live birth) (Cummings, 2009). Further, the OR is non-collapsible, making it

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

challenging to compare estimates between different studies (Richardson et al., 2017). Given that this is a prospective cohort study in which the exposure is measured prior to the outcome, it is more appropriate to directly estimate the relative risk of live birth. While it may seem strange to estimate the 'risk' of a good outcome, it should simply be thought of as the probability of having a live birth for a given woman during a single IVF cycle. However, while the relative risk is more appropriate than the OR, it can be harder to achieve convergence with the log-binomial model due to estimation under a constrained parameter space (Williamson et al., 2013).

We then considered three approaches which incorporate multiple IVF cycles per woman. These included a log-binomial mixed effects model, a GEE log-binomial model, and a CWGEE model. Clustering arises when multiple cycles per woman are analyzed, such that each woman contributes a cluster and each IVF cycle is a cluster member. Due to underlying biological differences between individuals, cycle outcomes are likely to be correlated within a cluster. We considered a log-binomial mixed effects model including a subject-specific effect to account for this *within-woman* correlation. Mixed effects models are useful for making subject-specific inferences in the context of unbalanced data, and address lack of independence by incorporating random effects in addition to fixed effects (Fitzmaurice et al., 2012; Yelland et al., 2011b). The subject-specific, or conditional, estimate may be interpreted as the relative risk of live birth, comparing two women with the same value of a random intercept but who differ in their age at cycle initiation. In situations in which the log-binomial model failed to converge, we fit a modified Poisson model including a subject-specific random intercept to estimate the association of maternal age at cycle initiation with the conditional probability of achieving a live birth. The modified Poisson approach is less prone to convergence problems and is a viable alternative to the log-binomial mixed effects model for estimating relative risks (Yelland et al., 2011a; Zou, 2004).

As an alternative, GEE models were fitted with a log link function and binomial distribution for evaluating age, and a Poisson distribution for evaluating ∑DEHP. These represent a marginal method: while mixed effects models estimate subject-specific effects, GEE models estimate population-averaged effects of exposure. Last of all, we considered a weighted GEE model to account for cluster size. Cluster size is considered to be informative, or non-ignorable, when the number of observations in a cluster is associated with the probability of the outcome. For IVF, the number of prior cycles that a patient has undergone is associated with her probability of success in the current cycle (Malizia et al., 2009). For example, couples with more severe infertility will likely undergo a greater number of IVF cycles before achieving a live birth, compared to couples with less severe infertility. When cluster size is informative, using an unweighted approach in marginal analyses will over-weight couples with the most severe infertility, leading to biased estimates. A weighted GEE approach, in which the weight is equal to the inverse of the cluster size, is not subject to this bias (Williamson et al., 2003; Huang and Leroux, 2011).

For all approaches, we conducted unadjusted and adjusted analyses. In the adjusted analyses, we included BMI (continuous, kg/m^2), maternal smoking history (ever, never), and SART infertility diagnosis (male, female, unexplained) as covariates. These covariates were chosen *a priori* for their known robust association with live birth following IVF treatment (Fedorcsák et al., 2004; Malizia et al., 2009; Vaegter et al., 2017). We did not include other predictors of live birth because the study population was quite homogenous; characteristics such as race and ethnicity were not likely to cause significant confounding. In addition, we did not include prior treatment in the first-cycle models because these were intended to demonstrate how a model that did *not* include information on prior treatment or number of cycles would behave (in contrast to the multiple-cycle models). In the analysis of ∑DEHP, we further adjusted for maternal age at cycle initiation (continuous) and age squared (continuous). For all models, we tested for nonlinearity in the association between each of the continuous predictors and live birth by adding a quadratic term and comparing model fit with a likelihood ratio test. For ordinal covariates, we utilized the Akaike information criterion to compare models with the ordinal variable versus those with indicators of levels of the ordinal variable, where possible. Finally, we assessed trend across quartiles for each model evaluating age, using the median values of each quartile as a continuous variable in regression models. Statistical analyses were conducted using SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA). *P*-values <0.05 were considered significant.

To help inform our analyses and to better understand the correlation structure among multiple cycles, we conducted two supporting evaluations. First, we assessed the association between cycle number and probability of live birth, in order to verify our assumption of informative cluster size and to consider the utility of the CWGEE approach. Second, we evaluated the working correlation matrix estimated by the unadjusted, unweighted GEE model under both the unstructured and the compound symmetry covariance structures. Consideration of the correlation structure is an important component of statistical analysis and can help elucidate underlying relationships within the data.

Results

Our analysis included 442 women who were enrolled in the EARTH study between November 2004 and June 2017 and who initiated at least one IVF cycle. At study entry, the participants were on average 35 years old (SD = 4), with a mean BMI of 24 kg/m^2 (SD = 4) (Table I). Thirty-three percent of study participants initiated multiple cycles (range: 1–5 cycles) and there was a total of 642 cycles for the 442 women. Overall, 37% of all cycles resulted in a live birth. Women in the first, second, third, and fourth quartile of age were 21–32, 33–35, 36–38 and 39–43 years of age, respectively. Among the 401 women with ∑DEHP measures (n = 575 cycles), the median urinary concentration of the per-cycle specific gravity-adjusted ∑DEHP was 0.13 μg/l (interquartile range (IQR): 0.07, 0.26), and the maximum value was 5.23 μg/l. The ∑DEHP concentration ranges of each quartile were: 0.010–0.065 μg/l, 0.066–0.123 μg/l, 0.124–0.282 μg/l, and 0.285–5.229 μg/l, respectively. There were no appreciable differences between the characteristics of women with and without ∑DEHP measurements (Table I).

Associations of age with live birth: models restricted to the first cycle

We observed significant associations between maternal age and live birth in all models. On average in the first IVF cycle, the probability of live birth was 55% lower among women in the oldest age group, compared to those in the youngest age group (RR = 0.45, 95% CI: 0.32, 0.64) (Table II). This estimate was slightly attenuated after adjusting for BMI, smoking status and infertility diagnosis (adjusted RR (aRR) = 0.49, 95% CI: 0.35, 0.70). For each quartile of age, the OR was farther from the null compared to the RR estimated by the log-binomial model, as expected based on their different interpretation. For example, the odds of live birth were 70% lower among the oldest women compared to the youngest, adjusting for BMI, smoking status, and infertility diagnosis (aOR = 0.30, 95% CI: 0.17, 0.53).

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

Table II Unadjusted and adjusted associations between maternal age at cycle initiation (years, in quartiles) and live birth (n = 442 women, 642 cycles).

	First Cycle Models		*Multiple Cycle Models*		
	Logistic OR (95% CI)	**Log Binomial RR (95% CI)**	**Mixed Effects RR (95% CI)**	**GEE RR (95% CI)**	**CWGEE RR (95% CI)**
Model 1: Unadjusted					
Q1	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
Q2	0.68 (0.40, 1.14)	0.83 (0.65, 1.07)	0.89 (0.64, 1.24)	0.91 (0.71, 1.16)	0.87 (0.70, 1.08)
Q3	0.48 (0.28, 0.82)	0.68 (0.51, 0.91)	0.65 (0.45, 0.93)	0.66 (0.50, 0.87)	0.68 (0.52, 0.89)
Q4	0.26 (0.15, 0.45)	0.45 (0.32, 0.64)	0.46 (0.31, 0.67)	0.47 (0.34, 0.64)	0.48 (0.35, 0.65)
Model 2: Adjusted for BMI, smoking status and infertility diagnosis					
Q1	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)	1.00 (ref)
Q2	0.72 (0.42, 1.21)	0.84 (0.66, 1.07)	0.90 (0.65, 1.26)	0.91 (0.73, 1.14)	0.89 (0.73, 1.08)
Q3	0.52 (0.30, 0.90)	0.71 (0.53, 0.95)	0.67 (0.47, 0.96)	0.66 (0.50, 0.88)	0.71 (0.55, 0.92)
Q4	0.30 (0.17, 0.53)	0.49 (0.35, 0.70)	0.51 (0.35, 0.75)	0.51 (0.37, 0.69)	0.53 (0.39, 0.72)

GEE, generalized estimating equation; CWGEE, cluster-weighted generalized estimating equation; OR, odds ratio; RR, risk ratio.
Age Quartiles: Q1, 21–32; Q2, 33–35; Q3, 36–38; Q4, 39–43 years.

Association of age with live birth: models including multiple cycles per woman

The probability of live birth for an individual in the oldest age category was 54% lower compared to an individual in the youngest age category, as estimated by the unadjusted mixed effects model (RR = 0.46, 95% CI: 0.31, 0.67) (Table II). After adjusting for BMI, smoking status and infertility diagnosis, the relative risk was 0.51 (95% CI: 0.35, 0.75). Results obtained from the GEE models were similar to those estimated by the mixed effects models. The unadjusted, unweighted GEE model yielded an average probability of live birth that was 53% lower among the oldest group of women compared to the youngest (RR = 0.47, 95% CI: 0.34, 0.64) (Table II). Similarly, the CWGEE yielded an unadjusted marginal relative risk of 0.48 (95% CI: 0.35, 0.65). After adjusting for BMI, smoking status, and infertility diagnosis, the relative risks were slightly attenuated for both the GEE and CWGEE models.

In general, the CIs were widest for the mixed effects models and narrowest for the GEE and CWGEE models. The 95% CIs tended to be narrower for the CWGEE compared to the GEE models. For example, the CI widths for comparing the oldest and youngest age groups, obtained with the fully adjusted mixed effects, GEE, and CWGEE models, were 0.44, 0.42 and 0.33, respectively (Table II). Finally, increased quartiles of maternal age were associated with a downward trend in the probability of live birth. This trend was significant ($P < 0.005$) across all models (data not shown).

Association of ∑DEHP with live birth

Among our overall cohort (n = 442 women), 401 women undergoing 575 cycles had complete information on urinary concentrations of DEHP metabolites. We did not observe any association of urinary ∑DEHP metabolite concentrations with the probability of live birth in this cohort (Table III), either in models restricted to the first cycle or accounting for multiple cycles. However, the estimates were surprisingly different between the first-cycle and multiple-cycle models. For example, the adjusted RR comparing women with the highest and lowest quartiles of ∑DEHP for each model incorporating multiple cycles was <1 (0.87, 0.87 and 0.95 for mixed effect, GEE and CWGEE, respectively), while that for the log-binomial model restricted to the first cycle was 1.24. While neither setting yielded statistically significant associations, restricting to the first cycle could lead investigators to interpret their findings quite differently.

Evaluating correlation structure among live birth outcomes in multiple IVF cycles

After adjusting for maternal age at cycle initiation, BMI, smoking status and infertility diagnosis, each additional IVF cycle was associated with a 32% decrease in the probability of achieving a live birth (aRR = 0.68, 95% CI: 0.55, 0.84), suggesting that cluster size is informative (data not shown). The working correlation matrix estimated by the unadjusted, unweighted GEE model, specifying an unstructured covariance, yielded

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

Table III Unadjusted and adjusted associations between ∑DEHP (quartiles) and live birth (n = 401 women, 575 cycles).

	First Cycle Models		*Multiple Cycle Models*		
	Logistic OR (95% CI)	**Log Binomial RR (95% CI)**	**Mixed Effects RR (95% CI)**	**GEE RR (95% CI)**	**CWGEE RR (95% CI)**
Model 1: Unadjusted					
Q1	1.00	1.00	1.00	1.00	1.00
	(ref)	(ref)	(ref)	(ref)	(ref)
Q2	1.42	1.22	1.12	1.11	1.07
	(0.83, 2.43)	(0.90, 1.65)	(0.78, 1.63)	(0.84, 1.47)	(0.82, 1.39)
Q3	1.14	1.08	0.96	0.99	0.99
	(0.67, 1.94)	(0.79, 1.48)	(0.66, 1.40)	(0.73, 1.34)	(0.76, 1.31)
Q4	1.29	1.16	0.82	0.90	0.92
	(0.72, 2.33)	(0.83, 1.63)	(0.55, 1.24)	(0.66, 1.24)	(0.68, 1.24)
Model 2: Adjusted for age, age^2, BMI, smoking status and infertility diagnosis					
Q1	1.00	1.00	1.00	1.00	1.00
	(ref)	(ref)	(ref)	(ref)	(ref)
Q2	1.34	1.10	1.07	1.08	1.03
	(0.76, 2.36)	(0.83, 1.46)	(0.74, 1.55)	(0.82, 1.40)	(0.80, 1.32)
Q3	1.17	1.09	0.98	0.98	1.01
	(0.67, 2.05)	(0.81, 1.47)	(0.67, 1.43)	(0.75, 1.29)	(0.78, 1.30)
Q4	1.36	1.24	0.87	0.87	0.95
	(0.73, 2.56)	(0.90, 1.71)	(0.58, 1.32)	(0.63, 1.20)	(0.72, 1.26)

∑DEHP Quartiles: Q1, 0.010–0.065 μg/l; Q2, 0.066–0.123 μg/l; Q3, 0.124–0.282 μg/l; Q4, 0.285–5.229 μg/l.

Table IV Model characteristics.

Model	Properties	SAS procedure[a]	Interpretation
Logistic Regression	• Requires independent observations	PROC GENMOD	
Log-Binomial Regression	• Requires independent observations	PROC GENMOD	
Mixed Effects Models	• Accounts for correlated observations • Can yield unbiased estimates with unbalanced cluster sizes given adjustment for covariates predicting imbalance	PROC GLIMMIX	Conditional
Generalized Estimating Equation Models	• Accounts for correlated observations • Does not require distributional assumptions • Assumes imbalance in data follows MCAR structure	PROC GENMOD	Marginal
Cluster-Weighted Generalized Estimating Equation Models	• Accounts for correlated observations • Does not require distributional assumptions • Accounts for non-ignorable cluster size	PROC GENMOD	Marginal

MCAR, Missing completely at random.
[a]SAS statistical procedure (PROC) for generalized linear models (GENMOD) or generalized linear mixed models (GLIMMIX).

negative correlations between the outcomes of early cycles and positive correlations between the outcomes of later cycles. This could occur if the couples who go on to fourth and fifth cycles have such severe infertility that they never achieve a live birth. In contrast, the compound symmetry covariance structure assumes that all cycles within an individual are equally correlated with one another. It is frequently utilized in studies of IVF, as it requires estimation of only two covariance parameters. However, given the results above, the compound symmetry structure does not seem appropriate. Indeed, the GEE models did not converge when we applied the compound symmetry covariance structure. Furthermore, a mixed effects model with a random intercept and no additional covariance structure is equivalent to specifying the compound symmetry covariance and may not be appropriate.

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

Discussion

In this cohort of subfertile couples undergoing ART, we examined several approaches to analyzing IVF data with multiple cycles and a non-rare, binary outcome. Specifically, we considered log-binomial and logistic regression models restricting to the first cycle, as well as mixed effects models, GEE models and CWGEE models when all cycles were included (summarized in Table IV). While several models performed adequately and yielded overall similar results, CWGEE models generally yielded the narrowest CIs. This was unexpected, as the CWGEE model includes fractional cycles summing to one per woman, rather than each woman contributing the total number of her cycles to the sample size. The cluster-weighted approach is a computationally simple way of addressing non-ignorable (informative) cluster size. In sensitivity analyses, consistent with previous studies, we observed that the number of IVF cycles a woman has undergone is a significant predictor of live birth (Malizia et al., 2009). Specifically, each additional IVF cycle is associated with a reduced probability of live birth. Thus, investigators should treat cluster size as informative in the IVF setting. Further, GEE models are generally robust against mis-specification of the underlying covariance structure (Fitzmaurice et al., 2012). Finally, the population-averaged interpretation of the CWGEE parameter estimate is clinically relevant. Specifically, the RR compares the probability of live birth among cycles of women in the fourth quartile of age to cycles of women in the first quartile of age. In contrast, mixed effects models estimate subject-specific associations, which can be somewhat less intuitive and more complicated to interpret. The subject-specific, or conditional, RR compares the probability of live birth within a woman, between two cycles that differ in quartile of age. This interpretation becomes more complex when covariate values are constant within a cluster. While differences in interpretation between the marginal (population-averaged) and conditional (subject-specific) models are nuanced, they should be considered when developing analysis plans and inferences. Specifically, investigators should compare the relevance of population versus individual level inferences in the context of the study aims.

One limitation of the CWGEE is that it may perform poorly for covariates with within-cluster variation (Huang and Leroux, 2011). Moving forward, more sophisticated statistical techniques such as the type-3 Doubly Weighted GEE (DWGEE3) estimator should be explored to handle cluster-varying exposures (Huang and Leroux, 2011). Further, G-Methods should be explored for handling time-varying confounding in scenarios where covariates vary by cycle and are affected by past exposure (Robins, 1986). An exploration of these techniques was beyond the scope of the present investigation.

In this investigation, the probability of live birth was around 50% lower among the oldest group of women compared to the youngest. However, contrary to our hypothesis, we observed no associations between ∑DEHP and live birth. Given previous findings (Hauser et al., 2016), our results may be attributed to the evolving composition of this cohort. The previous investigation included EARTH Study participants who were recruited between November 2004 and April 2012, totaling 256 women and 375 IVF cycles. We analyzed data from 401 women undergoing 575 cycles, between November 2004 and June 2017. We are encouraged to find that these additional years (April 2012–June 2017) correspond to declining urinary concentrations of DEHP metabolites. Specifically, the median urinary concentration of the per-cycle specific gravity-adjusted ∑DEHP was 0.19 μg/l (IQR: 0.10, 0.42) for 2004–2012, 0.07 μg/l (IQR: 0.05, 0.11) for 2012–2017, and 0.13 μg/l (IQR: 0.07, 0.26) for 2004–2017 overall. These changes reflect declining exposure to DEHP in this cohort and have likely attenuated the association between urinary ∑DEHP metabolites and live birth.

To the best of our knowledge, only one study has compared multiple methods for analyzing IVF data with multiple cycles (Missmer *et al.*, 2011): however, there were several limitations to this investigation. First, the authors exclusively estimated ORs for live birth. In IVF settings, the OR may not approximate the RR, as the probability of a live birth in the first cycle is around 25–30% (Cummings, 2009; Malizia et al., 2009; Pearson *et al.*, 2009). Notably, the estimated OR will be farther from the null than the corresponding RR, even though the strength of the association is the same. This discrepancy may lead investigators to conclude that the association between an exposure and outcome is stronger than it actually is. Second, the authors acknowledged that GEE models perform poorly with non-ignorable cluster size but did not address this further. We have gone beyond this and demonstrated the CWGEE approach, which accounts for informative cluster size. Finally, the authors restricted their analysis to cycles with an embryo transfer. We caution against this restriction, as it may lead to bias when the exposure of interest is associated with IVF failure points that occur prior to embryo transfer (Messerlian and Gaskins, 2017).

Among studies of live birth conducted within the EARTH cohort, it is standard practice to include multiple IVF cycles per woman in statistical analysis. ORs are often estimated with logistic mixed effects models including a random intercept. The least squares means of the fixed effects are then computed and the predicted marginal probabilities are presented for the average value of all covariates (Searle *et al.*, 1980; Hauser et al., 2016; Mínguez-Alarcón *et al.*, 2016; Gaskins *et al.*, 2018). To demonstrate this, we have computed the probability of live birth for each quartile of age with the least-squares means approach for the mixed effects models (Supplementary Table SI, SAS Code in Supplementary Data). It is also possible to estimate the means at specific levels of each covariate. Presenting marginal probabilities alongside **relative risks** can improve the interpretability of study results. However, we suggest the inclusion of the CWGEE approach to estimate **relative risks** for binary outcomes.

In addition to methods employed within the EARTH Study, a variety of statistical approaches have been considered in the literature that address the complex structure of IVF data. Among studies that include multiple cycles per woman, discrete survival analysis has been employed to consider the association of an exposure with the number of cycles until live birth or with the cumulative rate of live birth (Malizia et al., 2009). Time-to-event methods have also been expanded to consider multiple points of failure within an individual (Maity *et al.*, 2014). In addition to survival analysis, the embryo-uterus (EU) approach models the probability of live birth as a function of embryo viability and maternal characteristics (Speirs *et al.*, 1983). This method has been demonstrated and expanded to account for multiple cycles, although it is not widely used (Baeten *et al.*, 1993; Zhou and Weinberg, 1998; Dukic and Hogan, 2002; Missmer et al., 2011).

The present analysis had two potential limitations. First, we did not address differential loss to follow-up in our statistical approaches. In particular, couples who discontinue treatment after a failed cycle may

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

have different exposure characteristics than couples who undergo a subsequent cycle. For example, advanced maternal age and poor prognosis has been associated with an increased rate of treatment discontinuation (Dodge *et al.*, 2017). Alternatively, many couples discontinue treatment because they have conceived spontaneously (Domar *et al.*, 2018). Unfortunately, it is not always possible to distinguish couples who discontinue treatment from those who exit the study or are simply between cycles. Within these constraints, inverse probability weighting has been explored to handle differential patterns in treatment discontinuation (Robins *et al.*, 2000; Modest *et al.*, 2018). Second, this analysis was limited to an example dataset, in which the true effect of any exposure is unknown. While this allows us to observe model performance in the context of real data, future analyses should be conducted in simulation studies under a range of plausible assumptions to further evaluate the appropriateness of each approach. Simulation studies enable investigators to determine whether CIs have achieved the nominal level of coverage, to evaluate the bias of an estimator, and to better understand the relative precision of different approaches.

Conclusions

There are two primary methodological challenges when incorporating multiple cycles in the analysis of IVF data. First, failure to account for correlated outcomes within individuals may lead to invalid CIs. Second, failure to account for informative cluster size may lead to biased estimates. Implementing appropriate statistical methods in studies of IVF is important for three reasons. First, couples presenting for treatment should be counseled with unbiased estimates of the probability of live birth, either in any given cycle or over the course of treatment. Second, there is a growing body of research aimed at advancing IVF techniques and success rates, and the use of appropriate methodology can facilitate the improvement of clinical outcomes in infertility care. Third, IVF provides a unique opportunity to study the underlying causes and modifiable risk factors of infertility. It is critical to ask appropriate questions and employ best practices in order to make inferences that could further our understanding of these processes. However, there is disagreement among investigators on whether to present odds or risks, and how to best incorporate data from multiple cycles (Messerlian and Gaskins, 2017). We have evaluated several methods to address these concerns and prefer the CWGEE model. Further, we stress the importance of estimating risks rather than odds. In general, analysis decisions should be made *a priori*, and should be guided by the underlying structure of one's data and target of inference.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Acknowledgements

The authors gratefully acknowledge all members of the EARTH study team, specifically the Harvard T.H. Chan School of Public Health research staff Jennifer Ford, Myra Keller and Ramace Dadd, physicians and staff at Massachusetts General Hospital fertility center. A special thank you to all of the study participants.

Authors' roles

All authors of this article have made substantial contributions to the development of this work, from its conception and design, through data acquisition and analysis, interpretation of results, and drafting and revision for intellectual content. Each author has approved the final version to be published.

Funding

Grants (R01ES022955, R01ES009718 and P30ES000002) from the National Institutes of Health.

Conflict of interest

None of the authors has any conflicts of interest to declare.

References

ASRM. SART Data Release: 2015 Preliminary and 2014 Final. 2017, ASRM Press Release and Bulletin Volume 19, Number 15.

Baeten S, Bouckaert A, Loumaye E, Thomas K. A regression model for the rate of success of in vitro fertilization. *Stat Med* 1993;**12**:1543–1553.

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.

Centers for Disease Control and Prevention. Fourth National Report on Human Exposure to Environmental Chemicals, Updated Tables, February 2015. 2013.

Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. 2013 Assisted Reproductive Technology Fertility Clinic Success Rates Report. 2013. US Dept of Health and Human Services.

Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. 2015 Assisted Reproductive Technology Fertility Clinic Success Rates Report. 2017. US Dept of Health and Human Services, Atlanta.

Cummings P. The relative merits of risk ratios and odds ratios. *Arch Pediatr Adolesc Med* 2009;**163**:438–445.

Dodge LE, Sakkas D, Hacker MR, Feuerstein R, Domar AD. The impact of younger age on treatment discontinuation in insured IVF patients. *J Assist Reprod Genet* 2017;**34**:209–215.

Domar AD, Rooney K, Hacker MR, Sakkas D, Dodge LE. Burden of care is the primary reason why insured women terminate in vitro fertilization treatment. *Fertil Steril* 2018;**109**:1121–1126.

Dukic V, Hogan JW. A hierarchical Bayesian approach to modeling embryo implantation following in vitro fertilization. *Biostatistics* 2002;**3**:361–377.

Fedorcsák P, Dale PO, Storeng R, Ertzeid G, Bjercke S, Oldereid N, Omland AK, Åbyholm T, Tanbo T. Impact of overweight and underweight on assisted reproduction treatment. *Hum Reprod* 2004;**19**:2523–2528.

Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley & Sons, 2012.

Gaskins AJ, Hart JE, Mínguez-Alarcón L, Chavarro JE, Laden F, Coull BA, Ford JB, Souter I, Hauser R. Residential proximity to major roadways and traffic in relation to outcomes of in vitro fertilization. *Environ Int* 2018;**115**:239–246.

Hauser R, Gaskins AJ, Souter I, Smith KW, Dodge LE, Ehrlich S, Meeker JD, Calafat AM, Williams PL, Team ES. Urinary phthalate metabolite concentrations and reproductive outcomes among women undergoing

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022

in vitro fertilization: results from the EARTH study. *Environ Health Perspect* 2016;**124**:831.

Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg* 1990;**5**:46–51.

Huang Y, Leroux B. Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics* 2011;**67**:843–851.

Louis JF, Thoma ME, Sørensen DN, McLain AC, King RB, Sundaram R, Keiding N, Buck Louis GM. The prevalence of couple infertility in the United States from a male perspective: evidence from a nationally representative sample. *Andrology* 2013;**1**:741–748.

Maity A, Williams PL, Ryan L, Missmer SA, Coull BA, Hauser R. Analysis of in vitro fertilization data with multiple outcomes using discrete time-to-event analysis. *Stat Med* 2014;**33**:1738–1749.

Malizia BA, Hacker MR, Penzias AS. Cumulative live-birth rates after in vitro fertilization. *N Engl J Med* 2009;**360**:236–243.

Messerlian C, Gaskins AJ. Epidemiologic approaches for studying assisted reproductive technologies: design, methods, analysis, and interpretation. *Curr Epidemiol Rep* 2017;**4**:124–132.

Messerlian C, Williams PL, Ford JB, Chavarro JE, Mínguez-Alarcón L, Dadd R, Braun JM, Gaskins AJ, Meeker JD, James-Todd T. The Environment and Reproductive Health (EARTH) Study: a prospective preconception cohort. *Hum Reprod Open* 2018;**2018**:hoy001.

Missmer SA, Pearson KR, Ryan LM, Meeker JD, Cramer DW, Hauser R. Analysis of multiple-cycle data from couples undergoing in vitro fertilization: methodologic issues and statistical approaches. *Epidemiology* 2011;**22**:497–504.

Modest AM, Wise LA, Fox MP, Weuve J, Penzias AS, Hacker MR. IVF success corrected for drop-out: use of inverse probability weighting. *Hum Reprod* 2018;**33**:2295–2301.

Mínguez-Alarcón L, Chiu Y-H, Messerlian C, Williams PL, Sabatini ME, Toth TL, Ford JB, Calafat AM, Hauser R. Urinary paraben concentrations and in vitro fertilization outcomes among women from a fertility clinic. *Fertil Steril* 2016;**105**:714–721.

Pearson KR, Hauser R, Cramer DW, Missmer SA. Point of failure as a predictor of in vitro fertilization treatment discontinuation. *Fertil Steril* 2009;**91**:1483–1485.

Richardson TS, Robins JM, Wang L. On modeling and estimation for the relative risk and risk difference. *J Am Stat Assoc* 2017;**112**:1121–1130.

Rittenberg V, Seshadri S, Sunkara SK, Sobaleva S, Oteng-Ntim E, El-Toukhy T. Effect of body mass index on IVF treatment outcome: an updated systematic review and meta-analysis. *Reprod Biomed Online* 2011;**23**:421–439.

Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model* 1986;**7**:1393–1512.

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;**11**:550–560.

Searle SR, Speed FM, Milliken GA. Population marginal means in the linear model: an alternative to least squares means. *Am Stat* 1980;**34**:216–221.

Silva MJ, Samandar E, Preau JL Jr, Reidy JA, Needham LL, Calafat AM. Quantification of 22 phthalate metabolites in human urine. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007;**860**:106–112.

Society for Assisted Reproductive Technology, American Society for Reproductive Medicine. Assisted reproductive technology in the United States: 2001 results generated from the American Society for Reproductive Medicine/Society for Assisted Reproductive Technology registry. *Fertil Steril* 2007;**87**:1253–1266.

Speirs AL, Lopata A, Gronow MJ, Kellow GN, Johnston WIH. Analysis of the benefits and risks of multiple embryo transfer. *Fertil Steril* 1983;**39**:468–471.

Thoma ME, McLain AC, Louis JF, King RB, Trumble AC, Sundaram R, Louis GMB. Prevalence of infertility in the United States as estimated by the current duration approach and a traditional constructed approach. *Fertil Steril* 2013;**99**:1324–1331.e1321.

Vaegter KK, Lakic TG, Olovsson M, Berglund L, Brodin T, Holte J. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers. *Fertil Steril* 2017;**107**:641–648.e642.

Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics* 2003;**59**:36–42.

Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. *Emerg Themes Epidemiol* 2013;**10**:14.

Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for estimating relative risks from clustered prospective data. *Am J Epidemiol* 2011a;**174**:984–992.

Yelland LN, Salter AB, Ryan P, Makrides M. Analysis of binary outcomes from randomised trials including multiple births: when should clustering be taken into account? *Paediatr Perinat Epidemiol* 2011b;**25**:283–297.

Zhou H, Weinberg CR. Evaluating effects of exposures on embryo viability and uterine receptivity in in vitro fertilization. *Stat Med* 1998;**17**:1601–1612.

Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol* 2004;**159**:702–706.

Downloaded from https://academic.oup.com/humrep/article/34/3/549/5255722 by Northeastern University Libraries user on 15 April 2022