Measuring and Modelling Multistage Treatment Outcomes: Method Development For In Vitro Fertilisation.

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy (PhD) in the Faculty of Biology, Medicine and Health

2017

Jack Wilkinson

Centre for Biostatistics

School of Health Sciences

This is a blank page

Contents

Word count: 78,666

List of Figures

List of Tables

List of abbreviations used in the thesis

AFC	Antral Follicle Count
AIC	Akaike's Information Criterion
AMH	Anti-mullerian Hormone
ASRM	American Society for Reproductive Medicine
ART	Assisted Reproductive Technologies
BESST	Birth Emphasising a Successful Singleton at Term
b-hCG	Beta Human Chorionic Gonadotropin
BMI	Body Mass Index
CI	Credible or Confidence Interval
COS	Controlled Ovarian Stimulation
COMET	Core Outcome Measures in Effectiveness Trials
COMMIT	Core Outcome Measures for Infertility and Priority Setting for Infertility
CROWN	Core Outcomes in Women's Health
DET	Double Embryo Transfer
DHEA	Dehydroepiandrosterone
eSET	Elective Single Embryo Transfer
ESHRE	European Society of Hum Reprod and Embryology
ET	Embryo Transfer
EU	Embryo-Uterine
FDA	Food and Drugs Agency
FET	Frozen Embryo Transfer
FSH	Follicle Stimulating Hormone
GEE	Generalised Estimating Equations
GnRH	-Releasing Hormone
hCG	Human Chorionic Gonadotropin
HFEA	Human Fertilisation and Embryology Authority

HMC	Hamiltonian Monte Carlo
HMG	Human Menopausal Gonadotropins
ICSI	Intracytoplasmic sperm injection
IMPRINT	Improving the Reporting of Clinical Trials of Infertility Treatments Statement
IQR	Interquartile Range
ISI	International Scientific Indexing
IUI	Intrauterine Insemination
IVF	In vitro fertilisation
LBE	Live Birth Event
LDR	Long Downregulation
LH	Lutenising Hormone
MAR	Missing at Random or Medically Assisted Reproduction (Journal Article 1)
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MNAR	Missing Not at Random
MI	Multiple Imputation
MOR	Metaphase II-Oocyte Output Rate
NICE	National Institute for Health and Care Excellence
NHS	National Health Service
OHSS	Ovarian Hyperstimulation Syndrome
OPU	Oocyte Pickup
RCTs	Randomised Controlled Trials
rFSH	Recombinant Follicle Stimulating Hormone
rhFSH	Recombinant Human Follicle Stimulating Hormone
ROC	Receiver Operating Characteristic
SART	Society for Assisted Reproductive Technologies
SART-CORS	Society for Assisted Reproductive Technologies Central Online Reporting System

SEM	Structural Equation Modelling
SET	Single Embryo Transfer
USOR	Ultrasound Guided Oocyte Recovery

ABSTRACT

In vitro fertilisation (IVF) comprises a sequence of interventions delivered to the treated woman and her embryos. Typically, IVF begins with a period of stimulation, where a patient's ovaries will be stimulated with drugs. This encourages the growth of follicles, which contain eggs. When the follicles are sufficiently developed, a trigger is given which causes eggs to be released. These are collected in a clinical procedure, and are fertilised with sperm, to produce embryos. The embryos are graded on the basis of morphological features, and an embryologist will select the best to be transferred to the patient's uterus. Once embryos are transferred, the hope is that the patient will have a successful pregnancy, culminating in the birth of one or more children.

The multistage treatment structure complicates measurement and modelling of IVF data. First, since patient responses to each of these interventions can be measured, outcome reporting is complicated by the sheer variety of outcome measures on offer. Second, since each intervention influences not only the immediate patient response, but also responses to interventions delivered subsequently, it is difficult to untangle the causal web underlying the IVF process. This in turn obfuscates the mechanisms by which IVF interventions ultimately influence the birth outcome, representing a barrier to the design of new treatment strategies. Routine statistical models are not capable of addressing this challenge. Bespoke approaches are required.

Our aims were to address methodological issues relating to the measurement and modelling of multistage IVF data. After reviewing the existing literature, we investigated outcome reporting practices on IVF clinic websites and randomised controlled trials (RCTs). This highlighted the multiplicity of measures in use. We identified 815 distinct outcomes in use in IVF RCTs and 51 on clinic websites. In relation to trials, this represents a barrier to both data synthesis and comparison between treatments. In relation to clinic websites, there is a concern that prospective patients will struggle to interpret the different measures, rendering truly informed decision-making impossible. Selective reporting is another inevitable consequence of outcome heterogeneity, common to both research and advertising of IVF. While recognising that different measures are suitable for different purposes, we argue for greater standardisation of outcome reporting. National reporting schemes offer one route to clear and consistent reporting for consumers.

Next, we adapted and extended joint modelling approaches used in econometrics, education, and toxicity research, to develop methods for the joint analysis of multistage outcomes. These methods can accommodate mixed response types (eg: count, ordinal, binary) measured at different levels of a multilevel data structure (eg: women and their embryos). We represented each response variable by a standard regression submodel, and linked these by specifying an underlying multivariate latent structure. Finding that this did not yield useful estimates of effects of upstream events on downstream responses, we extended the approach by introducing response variables as covariates in downstream submodels.

Throughout, we emphasise real datasets and research questions. We conclude by building a model to estimate the effects of ovarian stimulation on uterine receptivity, which is complicated by the fact that stimulation also contributes to the pool of embryos available for transfer. Our results suggest deleterious effects of stimulation on embryo implantation and live birth.

DECLARATION

I declare that S Table 18 and S Table 19, and some of the text describing the clinical protocols in Chapter 9 may have been included in a thesis submitted for the award of MD at the University of Manchester by Oybek Rustamov, the joint first author of Journal Article 4.

COPYRIGHT STATEMENT

Acknowledgements

First thanks and apologies go to my supervisors, Steve Roberts and Andy Vail. Thanks for giving me room to make useful mistakes and for asking tough questions. Apologies for starting fights with the UK IVF regulator, a Swiss drug company, and the IVF industry during the PhD. I hope we will continue to collaborate. To Andy again and Sarah Cotterill, thanks for giving me the best possible start to my statistical career and for setting me on the right track.

Thanks to my Centre for Biostatistics colleagues, who have provided moral support and valuable criticism throughout the past three years. In particular, thanks to my open plan colleagues past and present, including Matthias (is that how you spell it?), Mad Matt (this is a Mad Max joke), Antonia (for aesthetic critique, and unsuccessful transportation), Calvin J Heal (the embodiment of postmodern nomenclature), Sarah (for humiliating me in front of the leaders of our field), Liz (our logical limericist), and Martina (for enduring sustained melting). Thanks to Lesley, the youngest lecturer ever appointed, for acting as a personal mental health coach throughout the PhD, for keeping Sandbar in business, and for being a great pal. Thanks to all for pointing out my many errors.

Thanks to Cheryl and Daniel, for their expert advice. You have both been generous with your time and patient with my limited understanding of your specialty. Thanks also to Tope, Ann, Claudette, Greg, Keith, and the other staff at St Mary's who took the time to share their expertise.

Thanks to Pamela, and everyone else who was kind enough to share their experiences of subfertility with me. You have made the research better.

Thanks to my family. Stacy, for being patient and supportive, and for forcing me to apply to do a master's in statistics in the first place. That was a life-changing decision. And to Alexa, my bab-bab, thanks for making sure that I don't oversleep, and for reminding me what it's all for each day.

For Alexa

I. Motivation

Chapter 1. Introduction and literature review

Infertility affects around one in six heterosexual couples in the UK and while IVF is the treatment recommended by NICE, success rates are low (ranging from around 32% for women under 35 to less than 2% for women over 44, (NHS Choices, 2017)). IVF is a complex treatment with multiple stages, and patients may undergo the process several times before they are successful. As a result, treatment for infertility is often lengthy and both psychologically and financially burdensome, with many patients giving up without having a child (Roberts, et al., 2010a, Verberg, et al., 2008). The combination of high emotional stakes and low chances of success can make for a discouraging patient experience.

How can we make life better for people undergoing IVF? One way would be to reliably predict outcomes of IVF treatment, so that patients could be counselled regarding the likelihood of a successful outcome and how long it is likely to take to achieve it. A second way would be to increase that likelihood, by improving the treatment delivered. Both of these objectives require us to answer the same question: what should we consider to be a successful outcome of IVF? We can't predict success if we don't know what success looks like, and we can't improve treatment if we don't know the desired result. Even if there is broad agreement that the desired outcome of treatment is the birth of a healthy child, there remains considerable scope for disagreement between stakeholders: clinicians emphasise safety (eg: Min, et al., 2004); patients prefer expediency and a reduction in overall burden (Roberts, et al., 2010a); clinics have commercial interests and favour outcome definitions that cast their performance in a favourable light (Abdalla, et al., 2010). Since these perspectives may not coincide, there is a need to consider the implications of the choice of endpoint for the inferences we make about IVF.

1.1 IVF is a multistage treatment

IVF is a complex, multistage process consisting of stimulation of the ovaries, retrieval and fertilisation of eggs and the transfer of some or all of the resulting embryos to the patient's uterus (Van Voorhis, 2007). Some of these embryos may implant, resulting in pregnancy, and any of those implanted may result in a live birth. Those embryos not used for the initial ‘fresh’ transfer may be cryopreserved, so that they can later be thawed and transferred in subsequent attempts (Van Voorhis, 2007). A patient may repeat this entire sequence of stimulation followed by the transfer of fresh and then frozen embryos multiple times before achieving a live birth. Alternatively, a patient may abandon treatment without managing to have a child (Pelinck, et al., 2007, Soullier, et al., 2008, Verberg, et al., 2008, Verhagen, et al., 2008). Figure 1 shows the multistage and potentially cyclical nature of IVF treatment.

Due to the multistage nature of IVF, the treatment is heterogeneous in nature. Clinics differ in the specific techniques and algorithms they use at each stage of the process, including how a patient's ovaries are stimulated and how embryos are cultured, selected and frozen (whether or not these variations actually impact clinical outcomes is beyond the scope of the present discussion). They also differ in relation to `higher-level' policies including how they select which patients to treat and how many cycles they will offer before refusing to provide further treatment (Sharif and Afnan, 2003). In addition to differences between centres, treatment will vary between patients within the same centre. The stimulation protocol may be to some extent personalised on the basis of presumed predictive markers of ovarian response, for example (La Marca and Sunkara, 2014). Moreover, the treatment delivered is to some extent dynamic or reactive, with the outcome at earlier stages determining what is done subsequently. In particular, the number and maturity of eggs retrieved post-stimulation restricts the available options in the next stages of treatment. Small numbers or immaturity of eggs will result in fewer good embryos available for transfer, and this will dictate the number and quality of embryos transferred as well as the number of spare embryos available for frozen transfers. Conversely, if the patient has a hyper-response to stimulation, resulting in the release of an excessive yield of eggs, it might be necessary to freeze them all for later transfer as a precaution against ovarian hyper stimulation syndrome (OHSS) (Fiedler and Ezcurra, 2012). Part of the challenge of measuring IVF responses lies in understanding these sources of variation and working out how they should be accommodated in outcome measures and statistical analyses.

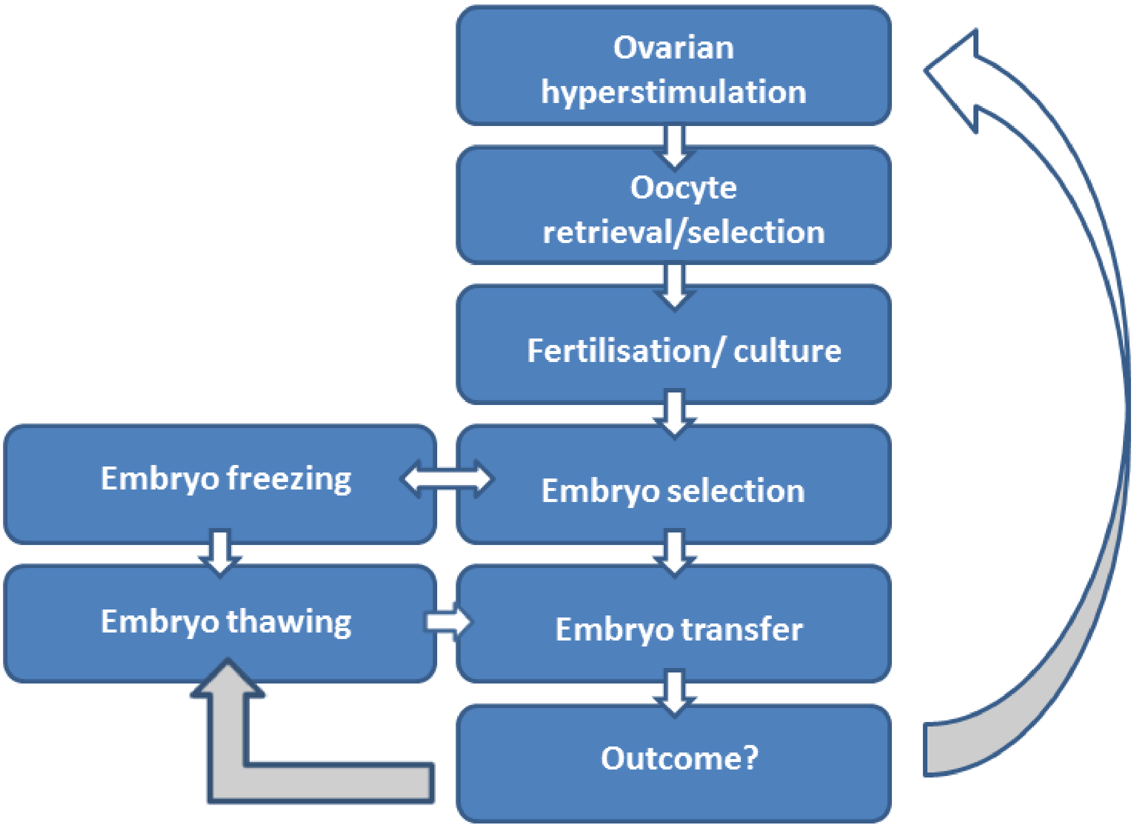


Figure 1: IVF as a multistage treatment. The ovaries are stimulated to produce oocytes (eggs), which are collected and fertilised to make embryos. The embryos are cultured for several days and the best are selected for transfer. The remainder may be frozen for future use. If the embryo transfer fails (or results in a pregnancy which is not carried to term) then some of the frozen embryos might be thawed and transferred. Alternatively (or sometimes additionally, following a failed frozen transfer attempt) the patient may undergo another round of ovarian stimulation, starting a new cycle.

1.2 Measuring the outcome of IVF. What does the literature say?

Couples undergo IVF to have a baby. However, the simplicity of this goal belies the complexity of measuring success. As a result, there is an extensive literature pertaining to outcome definition in IVF, cataloguing an array of discordant perspectives. Much of this was triggered by a proposal made by Min and colleagues (2004) that the success of a particular IVF treatment should be evaluated according to whether or not it resulted in the birth of a singleton (as opposed to twin or triplet), term gestation (as opposed to premature) baby. Their assertion was that performance of an IVF clinic should be evaluated by the proportion of treatments that, once initiated, resulted in a birth meeting these criteria. They called this measure ‘Birth Emphasising a Successful Singleton at Term’, or ‘BESST’. BESST serves as an example of some of the things we have to consider when choosing an outcome measure for IVF. The authors specified the event that they thought should be taken as the target of treatment (singleton, term birth - the numerator of the measure). They also specified which treatments should be included when calculating the proportion (all treatments initiated – the denominator of the measure). In the article, they also explained the thinking behind the measure: BESST was designed to encourage patient safety (for reasons discussed later, singleton births are considered to be safer than twin births). We will see that the numerator and denominator have been chosen with this purpose in mind. These three elements (purpose, numerator and denominator) must be considered whenever we construct an IVF outcome measure.

1.2.1. What is the outcome measure for?

IVF outcome measures are used for different purposes. They are used to evaluate interventions, to convey information to patients regarding the likelihood of success, and to describe the performance of IVF clinics. It is important to recognise that any particular outcome measure may be more or less appropriate for one of these purposes compared to the others. For example, outcomes which are appropriate for trials and observational treatment studies may not be well suited for the purposes of national assessment of clinic performance. The primary reason for this is that once the outcome for measuring clinic performance has been set, there is scope for clinics to game the system by being selective over the patients they treat and the manner in which they treat them (Bird, et al., 2005, Sharif and Afnan, 2003). In the UK, clinics compete for patients who are encouraged by the Human Fertilisation and Embryology Authority (HFEA) to consider performance when choosing a clinic. In light of this competition, clinics can be expected to tailor practices to some degree in order to give a flattering impression to patients. The definition of the national performance indicator therefore has scope to influence the treatment that is delivered, a consequence which must be anticipated when selecting the measure to be employed. This is distinct from the case of trials and some observational studies designed to estimate the effectiveness of treatment regimens, where the assumption would be that treatment is delivered in a standardised way.

An example of how the self-interest of clinics could be manipulated for patient benefit is given by Abdalla and colleagues (2010). Here, the authors discuss the outcome `live births per oocyte retrieval', where any live births resulting from all of the fresh and frozen transfers arising from a single episode of ovarian stimulation are counted (later we will describe this as a 'cumulative' measure). The use of this outcome measure would encourage clinics to make greater use of frozen transfers, because this would increase the number of live births (the numerator) without increasing the number of oocyte retrievals (the denominator). Since a fraction gets bigger if we increase the numerator while holding the denominator fixed, this would translate to a higher ranking for any clinic that played along. The benefit for patients would be safer treatment, since greater use of frozen embryos is considered less hazardous than transferring large numbers of fresh embryos or commencing another episode of ovarian hyperstimulation. BESST was intended to have a similar effect, by counting only singleton, not twin, births in the numerator (the latter becomes more likely when multiple embryos are transferred at once). Different outcome measures may also have differential utility for the purposes of providing prognostic information as opposed to evaluating clinic performance (Abdalla, et al., 2010) or for representing clinician as opposed to patient opinion (Min, et al., 2004, Roberts, et al., 2010a). The intended function and audience for an outcome measure must be central to any assessment of its merits.

1.2.2. **IVF outcomes as a numerator and a denominator**

As we saw in our discussion of BESST (section 1.2), IVF outcomes are typically expressed as a numerator, representing a count of some event of interest, and a denominator, representing the unit of analysis and providing the context for the measurement (Abdalla, et al., 2010, Heijnen, et al., 2004). In this framework, the choice of outcome reduces to the selection of an appropriate numerator and denominator, or equivalently, an appropriate clinical event and unit of observation.

1.2.3. **The choice of numerator**

The prevailing view is that the clinical event or numerator that should be used to evaluate IVF is a live birth (or live birth event, see below) owing to the fact that the birth of a baby is the goal of any initiated treatment (eg: Abdalla, et al., 2010, Garrido, et al., 2011, Germond, et al., 2004, Heijnen, et al., 2004, Maheshwari, et al., 2015, Min, et al., 2004, Moragianni and Penzias, 2010). In relation to trials, the recommendation that live birth should be reported in all infertility studies has been included in a recent extension to the CONSORT (Consolidated Standards of Reporting Trials) statement (Moher, et al., 2010), known as IMPRINT (Improving the Reporting of Clinical Trials of Infertility Treatments (Legro, et al., 2014). This position isn't universally held; Griesinger and colleagues (2004) argue that ongoing pregnancy is a more relevant measure for the purpose of treatment evaluation. They argue that, once an ongoing pregnancy has been established, a live birth is dependent on prenatal diagnosis, antenatal and obsetric care, which they consider to be distinct from the core treatment delivered by a clinic and which may represent a source of post-intervention confounding in comparative studies. Braakhekke and colleagues (2014) repudiate these premises but arrive at the same conclusion; they argue that ongoing pregnancy should be used on pragmatic grounds, precisely because the correlation with live birth is so strong. Other appeals for an emphasis on outcomes other than live birth have been motivated by an interest in treatment or clinic performance rather than providing prognostic information to patients regarding the likelihood of success. For example, (Pinborg, et al., 2004) express concern that live birth alone is insufficient to evaluate the various stages of treatment, and suggest that the number of oocytes and ongoing pregnancies should also be reported. These intermediate, or ‘procedural’, endpoints convey information about the success of the stimulation and laboratory phases of treatment, respectively.

We return to the question of the intended function of the measure. As far as the intended audience comprises prospective patients, the argument for live birth as the numerator of the IVF outcome remains strong. If a course of treatment does not result in a baby, then a patient will consider this to represent a failure, no matter how good a response was achieved at the stimulation and implantation stages. Similarly, a pregnancy isn’t a good outcome if it ends in miscarriage. There are no partial successes in IVF. To the extent that the outcome is intended to be prognostic (as might be used when counselling couples regarding their chance of success), the prediction that matters to patients relates to the likelihood that they will have a child. Similarly, the objectives of clinicians are to counsel patients about their chances of having a child through IVF and to treat patients in such a way that this goal is realised (with supplementary duties of minimising risk of harm and psychological burden). Whenever interest lies in measuring how effective an IVF clinic or treatment is therefore, live birth appears to be the mandatory choice of numerator.

We should make a distinction between questions relating to the effectiveness of IVF and questions relating to efficacy and mechanism however. While we acknowledge that these terms are used quite heterogeneously, for current purposes we consider effectiveness to relate to pragmatic questions about whether or not a treatment works in practice, while efficacy relates to explanatory questions about whether or not the treatment has the desired effect under controlled circumstances (Ernst and Pittler, 2006). We consider questions of efficacy to precede questions of effectiveness in the sense that they might illustrate how a treatment may work in principle or put a hypothesised mechanism of action to the test, but cannot generally replace a pragmatic evaluation of whether or not a treatment improves clinical outcomes when implemented in unselected populations and realistic settings. An exception might be cases where a treatment is clearly demonstrated not to be efficacious; if an intervention doesn't display the intended effect under experimental conditions, then the plausibility of it improving outcomes if implemented in clinical practice is low. Our understanding of these terms is more general than that described by Stewart and colleagues (2011a), who defined 'efficacy' as the birth rate after some predefined number of IVF attempts and 'effectiveness' as the probability of having a baby if a patient starts IVF. We consider this usage to be both unusual and restrictive. IVF is really a sequence of interventions, each of which could be probed for efficacy, and live birth may not be the best measure for these inquiries. This was the argument made by Pinborg and colleagues, for example, when they suggested that number of oocytes should be used to evaluate the ovarian stimulation stage of treatment (2004).

To the extent that efficacy is a topic of interest in IVF then (and, as we shall show in Journal Article 2 (Chapter 4), it certainly appears to be), we should acknowledge the fact that not all measurement in IVF needs to be targeted at patients, and does therefore not need to convey information that is directly relevant to that group. Accordingly, we should be willing to entertain the possibility that alternative outcome measures are more appropriate for evaluating mechanistic questions about IVF. This might include procedural responses relating to the different stages of treatment, such as number of oocytes or measures of embryo quality. We are not making the dubious point here that these procedural outcomes can obviate clinical endpoints such as live birth by acting as surrogate outcomes (eg: Sunkara, et al., 2011). Rather, we are open to the possibility that different questions might require different standards of success, and therefore different numerators.

The choice of denominator

Audience and purpose are equally important for the choice of denominator. As we saw was the case with the choice of numerator, it is generally possible to categorise proposals for the denominator according to an emphasis on clinic performance or on more patient-centred information. Proposals falling in the former category include measuring events per transfer performed (Davies, et al., 2004, Heijnen, et al., 2004), per embryo transferred (Abdalla, et al., 2010, Heijnen, et al., 2004) or per oocyte retrieval (Chetkowski, 2014, Davies, et al., 2004, Heijnen, et al., 2004). As an example, Abdalla et al (2010) endorsed live birth event per embryo transferred as a measure of IVF clinic performance. Denominators in this category exclude earlier stages of the treatment from consideration, thereby restricting focus to a segment of the IVF process and to the subgroup of patients who actually reached a certain point in the sequence. Using 'oocyte retrieval' as the denominator leaves out patients whose treatment was cancelled for poor stimulation response, while using 'transfer performed' or 'embryo transferred' leaves out any patients who didn't make it as far as the transfer procedure. Restricting the measurement to a portion of the process that is directly under the control of clinics may be appropriate for the purposes of assessing the competence of IVF centres in relation to particular components of treatment (Abdalla, et al., 2010). However, the exclusionary nature of these denominators make them useless for the purposes of providing prognostic information to patients and clinicians, as they essentially presuppose successful 'upstream' treatment responses. Accordingly, they offer limited information to patients prior to starting IVF, many of whom will not complete all of the stages of treatment due to outright failure early on (9% of treatments in the US in 2016 did not reach the egg collection, for example, Society for Assisted Reproductive Technologies, 2016). The desire to report more patient-oriented outcomes motivates a second group of proposals for the denominator. This includes proposals to count events per episode of ovarian stimulation started (Griesinger, et al., 2004, Min, et al., 2004, Schieve and Reynolds, 2004), per patient (Vail and Gardener, 2003), per course of treatment initiated (Daya, 2005, Heijnen, et al., 2004), per ‘full cycle’, including all fresh and frozen embryo transfers from one episode of stimulation (Veleva, et al., 2009), or per some defined period of time or number of treatment attempts (Gnoth, et al., 2011, Heijnen, et al., 2004, Malizia, et al., 2009). These suggestions move the beginning of the observation period to coincide with the start of treatment, thereby including all patients who actually begin IVF. By way of illustration, until recently HFEA reported live birth rates per cycle started, where the start of a ‘cycle’ coincides with the commencement of ovarian stimulation for fresh transfers (see Journal Article 3, Chapter 5). Consequently, the success rates reported by HFEA for fresh cycles included those patients for whom the cycle was cancelled (due to poor or excessive stimulation response, for example) as having failed treatments. These rates are clearly more informative to a patient prior to the start of treatment in relation to the likelihood that they will have a successful outcome.

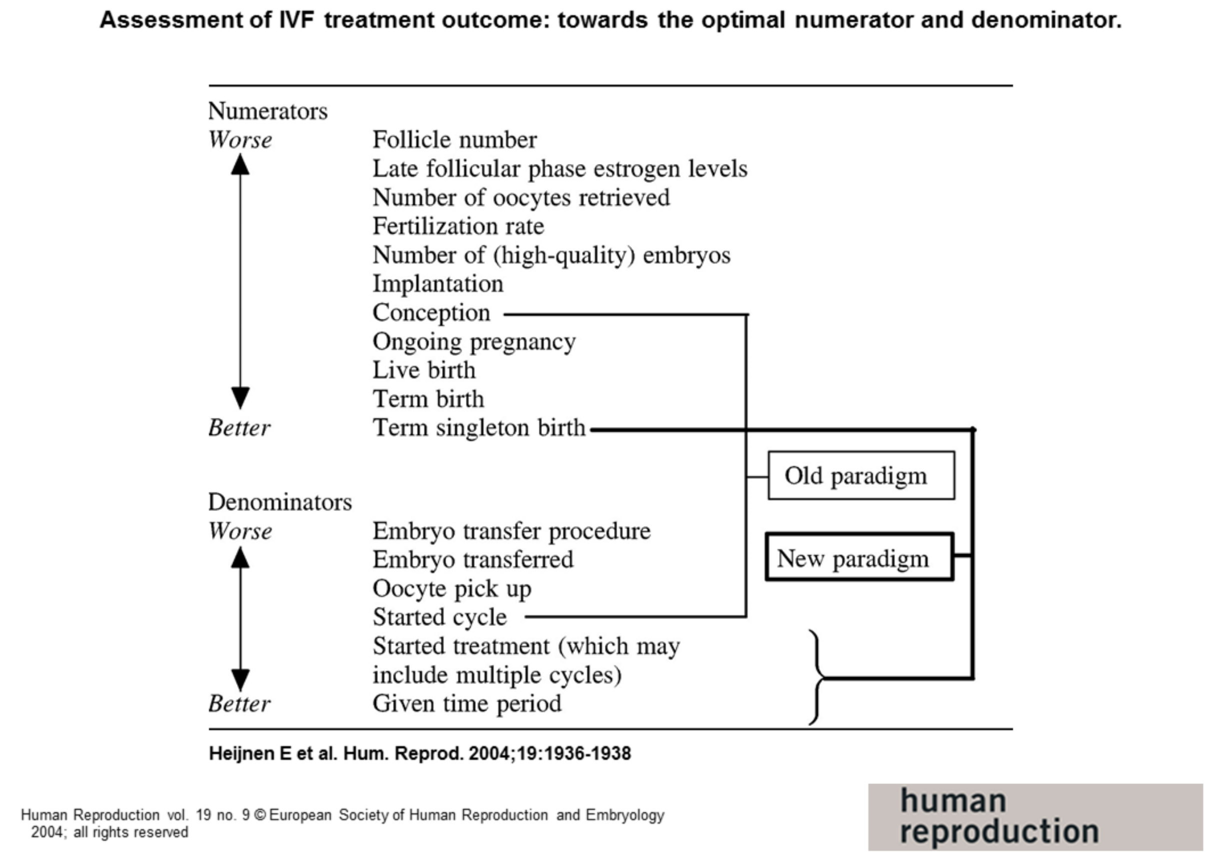
There is some concern that outcomes with inclusive denominators are susceptible to under-reporting of cancelled cycles by clinics in order to boost apparent performance (Abdalla, et al., 2010, Sharif and Afnan, 2003). Abdalla et al (2010) remark that substantial variation between centres and an increase in the number of centres in the UK reporting no or few cancelled cycles between 2002 and 2007 are suggestive of such behaviour. HFEA attempt to prevent this by requiring Intention to Treat forms to be completed upon commencement of treatment (Human Fertilisation and Embryology Authority, 2017). Even if these measures are successful, outcomes with denominators that include all patients starting treatment can still effectively be manipulated through patient selection (Sharif and Afnan, 2003). The fact that clinics may choose to select patients with reasonable or good prognosis means that patients with poor prognosis may be underrepresented in IVF datasets. This presents a concern about the generalisability of reported results to patients starting treatment.

Another difficulty relating to use of ‘per cycle started' as denominator is how to define the start of a ‘cycle’ for the transfer of frozen embryos. In the UK, HFEA consider a frozen transfer cycle to have started at the point of embryo thaw rather than at the point of drug administration, which occurs in the majority of these treatments. An implication of this is that cycles where stimulation was performed but the patient did not respond in a satisfactory manner (for example, the endometrium did not thicken sufficiently) are cancelled prior to thawing but are not reported as failed cycles. The implications of this are that some patients are excluded from the ‘per cycle started' figures for frozen transfers on the basis of prognosis. This is one reason why the rates for fresh and frozen transfers reported by the HFEA are not commensurable.

The remaining denominator options presented in the inclusive, patient-friendly category do not share this particular difficulty, because in addition to moving the start of the observation period to an earlier point in the IVF process, they also extend the scope of the observation period to cover full cycles (all fresh and frozen transfers from an episode of stimulation), or complete courses of treatment (which may include multiple stimulation episodes). Heijnen et al (2004) argued that measuring live birth over full courses of treatment (or in practice, over some defined extended period of time) is the most relevant way to evaluate IVF programmes (Figure 2, reproduced from Heijnen, et al., 2004). This reasoning motivates the use of ‘cumulative’ outcomes for IVF.

1.2.4. Cumulative outcomes over courses of IVF treatment

It can be argued that the most relevant piece of information for a patient about to start IVF is the probability that they will one day take home a child (Stewart, et al., 2011b). As previously noted, patients may undertake multiple attempts at treatment before having a child or giving up. To this end, some researchers have presented success rates ‘cumulatively’ over some maximum number of treatment attempts (Gnoth, et al., 2011, Luke, et al., 2012, Olivius, et al., 2002, Pelinck, et al., 2007, Soullier, et al., 2008, Stern, et al., 2010, Stewart, et al., 2011b, Sundstrom and Saldeen, 2009, Witsenburg, et al., 2005), (although several of these reported on pregnancy rather than birth outcomes). The term ‘cumulative’ may be misleading here, since it most usual for counting to cease at the first instance rather than to admit the possibility of subsequent births. However, given its prevalence in the literature, we adopt the term for the present discussion. Cumulative success rates share the property of having a denominator that exceeds a single cycle of treatment. However, there is variation in how cumulative outcomes are defined in relation to the number of treatment attempts included and the handling of cancelled treatments. For example, a 2010 review of cumulative live birth rates in IVF did not provide a definition of what the authors considered to constitute a cumulative live birth rate and included rates obtained from studies with heterogeneous designs (Moragianni and Penzias, 2010). This included a five-year retrospective cohort study of up to 10 cycles (mean (SD) = 2.3 (1.5) cycles) with nondonor oocytes, where cancelled cycles were included but patients who recieved treatment for less than 1 year were not (Malizia, et al., 2009), a two-year retrospective cohort study of up to 11 cycles (mean (SD) = 1.9 (1.2) cycles) which included donor oocyte transfers , where it was assumed that the first cycle recorded was the first cycle received (Stern, et al., 2010) and a systematic review of trials of elective single embryo transfer versus double embryo transfer where up to two cycles were considered (Gelbaya, et al., 2010). It is apparent that the cumulative rates will vary with the number of cycles considered and that it is not meaningful to compare rates from studies with different measures. The review authors do not appear to appreciate these points when they conclude that typical cumulative live birth rates following IVF are around 50%.



Assessment of IVF treatment outcome: towards the optimal numerator and denominator.

Numerators
Worse Follicle number
Late follicular phase estrogen levels
Number of oocytes retrieved
Fertilization rate
Number of (high-quality) embryos
Implantation
Conception
Ongoing pregnancy
Live birth
Term birth
Better Term singleton birth

Denominators
Worse Embryo transfer procedure
Embryo transferred
Oocyte pick up
Started cycle
Started treatment (which may include multiple cycles)
Better Given time period

Heijnen E et al. Hum. Reprod. 2004;19:1936-1938

Human Reproduction vol. 19 no. 9 © European Society of Human Reproduction and Embryology 2004; all rights reserved

human reproduction

Figure 2: Possible choices for numerator and denominator for IVF outcomes. Reproduced from Heijnen et al., 2004, pg 1937.

In practice, calculating cumulative outcomes presents methodological challenges induced by the fact that patients drop out for reasons that are unlikely to be independent of prognosis and the counterfactual outcome (that is to say, how things would have turned out had the patient continued treatment) (Daya, 2005). Existing approaches for dealing with this informative censoring are discussed below.

If interest lies in calculating rates over full courses of treatment, several practical challenges also exist. Firstly, given that some patients may undertake many cycles of treatment, the observation period has to be lengthy. If these rates are calculated prospectively, this means that there will be a delay before they can be reported. Secondly, a decision needs to be made regarding when to consider the observation period closed and to report the outcome. Given that a patient could in principle always return for further treatment, this decision will always be to some extent arbitrary and may not actually capture complete courses of IVF. Finally, patients may receive treatment at multiple centres. Given that most datasets come from single clinics, they do not contain information about treatment received prior to or following the cycles attempted at one centre. Although HFEA does share data on all treatment cycles performed in the UK, the present format does not allow for linkage of repeated cycles undertaken by a single patient across multiple centres. In the US, the feasibility of linking repeated cycles appearing in the national database in order to calculate cumulative rates has been demonstrated (Luke, et al., 2012, Stern, et al., 2010). The researchers made use of identifiable information to do this however (including name, date of birth and social security number), which isn't generally available to researchers.

As a result of these logistical and conceptual restrictions, it is unclear that developing general analytic models for outcomes following full courses of IVF treatment is a feasible or even coherent ambition. Patients may attend multiple centres and the decision to switch must be seen as informative (owing, for example, to a clinic's refusal to continue to treat patients with a certain number of failed attempts). Given the single-centre nature of most available IVF datasets (or the inability to follow patients between centres in collaborative multi-clinic datasets (Roberts, et al., 2010a), a more realistic ambition may lie in the development of analytic models for single-centre courses of IVF which can incorporate the informative censoring due to switching centres or dropping out completely. We anticipate that cumulative outcomes may be more relevant whenever we are interested in effectiveness rather than efficacy, since these measures are intended to capture patient-relevant event rates in realistic scenarios.

1.2.5. Multiple births and healthy babies

We've already encountered the proposal of Min and colleagues (2004) that the singleton, term gestation live birth rate per cycle started is the most relevant measure of the success of an IVF programme (the 'BESST' endpoint, discussed in 1.2). In endorsing BESST as the most suitable endpoint, the study authors were motivated by concerns over high rates of multiple births owing to the ubiquitous practice of transferring multiple embryos in each cycle. It is known that multiple births are associated with health complications for both the mother and children and the prevailing clinician opinion is that they should be avoided (Coetsier, et al., 2001, Stillman, et al., 2013). Accordingly, the authors argued that an appropriate outcome should penalise multiple births and reward the births of singleton babies. However, the BESST endpoint does not represent the perspective of many patients regarding what constitutes a successful result. Patients tend to emphasise the physical and psychological burden of treatment (Roberts, et al., 2010b). Consequently, many patients express a preference for a shorter treatment duration and for twins, removing the need to undergo further IVF treatment to have a second child (eg: Hojgaard, et al., 2007, Stillman, et al., 2013). This illustrates the point that it might not be possible to reconcile the diverse perspectives of stakeholders in a single outcome measure.

Even with increasing use of 'one at a time' elective single embryo transfer (eSET) as a measure to reduce multiple births, the question of how multiples should be handled in an outcome definition is likely to remain material for the foreseeable future; in the UK 51.3% of transfer cycles in 2014 involved two or three embryos and the multiple pregnancy rate (per pregnancy) was 15.9% (Human Fertilisation and Embryology Authority, 2016). In addition to proposals such as BESST that treat multiple births as failures, alternative possibilities include using 'live birth event' as the numerator of the outcome (thereby treating a singleton and multiple birth as equivalent), counting the number of babies born (thereby favouring multiples, reflecting common patient preferences but at the cost of encouraging unsafe stimulation and transfer practices) or scoring multiple birth events lower than singleton birth events, while still acknowledging this as a superior outcome to no birth (for example, scoring the birth of multiples as 0.5 compared to 1 for a singleton birth and 0 for no birth). The latter proposal runs contrary to our assertion that partial successes do not occur in IVF, although this is not a reason for dismissal in itself.

In addition to the classification of multiple births as treatment failures, the second controversial suggestion within BESST is that only live births following a term gestation should be counted as successes (Min, et al., 2004). The motivation for this suggestion is that the goal of IVF is to have a healthy baby, and so births of preterm babies, who might display low birthweight and otherwise have worse health outcomes, should be penalised. Arguments against this suggestion include the assertion that preterm birth may be outside of the control of the clinic and should not be counted against an IVF programme, and the observation that a preterm baby might be perfectly healthy (Griesinger, et al., 2004). Schieve & Reynolds (2004) are sympathetic to the idea of taking the health of the baby into account but question the utility of incorporating this information directly into an outcome measure owing to the fact that the risk of preterm birth is multifactorially determined. Instead, they suggest that singleton birth rates should be presented, stratified by risk factors for preterm birth (stratification in this way is discussed below). They also note that it would be desirable to consider early-life health information when determining whether a treatment is successful, but note the practical difficulties of linking information between different data sources and delayed outcome reporting.

1.2.6. Incorporating patient characteristics into outcome measures

Marginal live birth rates calculated on the basis of heterogeneous cohorts offer some information to a patient regarding their chances of success. These figures represent average or expected outcomes for patients undergoing IVF. However, success rates vary considerably between subgroups of patients defined on the basis of prognostic factors, and it is clearly more informative to present a patient with outcomes for similar patients undergoing treatment.

The approach formerly adopted by HFEA was to present results stratified by prognostic factors (age, infertility diagnosis and infertility duration), thereby allowing prospective patients to see success rates for people who are similar to them in relation to that particular factor. A patient's prognosis depends on multiple characteristics however, and patients within a particular stratum may be heterogeneous with respect to factors other than the stratification variable. Additionally, stratification requires the categorisation of continuous variables, so that some prognostic information is lost. Approaches based on multivariable models do not have these limitations. Using a modelling approach, predictions can be made on the basis of a multifactorial prognostic index and modelled relationships between outcome and predictor variables can be complex (Cai, et al., 2011, Nelson and Lawlor, 2011, Templeton, et al., 1996). The relative complexity of conducting and translating these analyses for lay audiences may prove to be prohibitive outside of a research context however.

An interesting question is how this prognosis can be updated on the basis of the outcomes of previous cycles or on the basis of intermediate outcomes of earlier stages within the same cycle. This will be discussed in the context of existing models in section 1.3

1.2.7. Should we be using time to event outcomes to measure IVF success?

We noted above that IVF treatment can be physically and psychologically burdensome, with many patients expressing a preference for twins so that there is no need for further treatment to have a second child (Roberts, et al., 2010b). Evidently, the overall probability of having a healthy baby is not the only consideration for a patient deciding whether or not to begin or continue treatment (Verberg, et al., 2008). An approach to outcome measurement that incorporates some of this additional information might be desirable. One possibility would be to present time to event outcomes for IVF. By presenting the expected time to success for patients with particular combinations of characteristics, patients and clinicians could consider not only the anticipated cumulative result of treatment but also the reality of what the journey is likely to entail.

There are several challenges that must be addressed in a time to event framework for IVF, some of which are general difficulties which arise when considering outcomes over multiple cycles of treatment and were discussed above in the context of cumulative outcomes. Most notably, the fact that censoring in IVF datasets is likely to be informative must be addressed. Additionally, there is the question of what scale should be used to measure the passage of time (Daya, 2005). One possibility would be to adopt a discrete time approach, whereby time is incremented by one unit with each initiated treatment attempt (or even at each stage within an attempt, (eg: Maity, et al., 2014). To the extent that treatment protocols vary however, this approach has the potential to be uninformative or misleading; providing an expected number of cycles needed does not convey how long this will actually take in a particular case. An approach based on real time might therefore be more appropriate (Daya, 2005).

Additional challenges relate to the issue of incorporating time-varying covariates and stage-specific intermediate outcomes across potentially multiple cycles when modelling the outcome. In relation to the outcome definition, in a time to event framework, it might be necessary to define the event of interest differently compared to those we have considered up to this point. It would not be appropriate to measure or report on time to live birth, for example, because this would favour shorter pregnancies and premature births. A more appropriate survival endpoint might be time to pregnancy leading to a live birth, or time to pregnancy leading to a term live birth.

Finally, the issue of how to emphasise patient and child safety with a time to event outcome must be considered. It would be perverse to reward hazardous treatment strategies including aggressive stimulation regimens and the transfer of large numbers of embryos because they achieve live birth in fewer attempts. As an example, if we use a time to event outcome and consider singleton and twin live birth events to be equivalent, double embryo transfer will be favoured over elective single embryo transfer; the latter strategy requires the subsequent transfer of frozen embryos to compensate for the reduced success rate in the initial cycle (McLernon, et al., 2010, Roberts, et al., 2010b). As noted above, this runs contrary to the concerns of many clinicians, and represents a considerable objection to the use of time to event outcomes in IVF.

1.2.8. A summary of section 1.2

A review of the literature relating to the choice of IVF outcome suggests that any measure of effectiveness should incorporate the probability of attaining a live birth while conditioning on patient characteristics in an appropriate and transparent way. Presenting outcomes over full courses of treatment would be of considerable value to prospective patients faced with the decision of whether or not to undergo treatment. However conceptual, methodological and practical difficulties lead to doubts about this as a realistic goal. Predicting outcomes over extended periods of time is likely to represent a more coherent endeavour. These extended courses of treatment can be emotionally and physically testing, and confer a certain degree of risk, with the implication that there are relevant considerations surplus to the question of how likely it is that a treatment will produce a baby. To this end, the importance of handling multiple birth events appropriately has been discussed (if not settled). A time to event framework has been suggested as an alternative mode of measuring success, although concerns over the implications of a 'quicker is better' approach to delivering treatment might rule this out.

We have also noted that the most appropriate outcome measure is likely to vary according to the intended audience and objective. In particular, we have drawn a distinction between questions of effectiveness and efficacy, and have noted that mechanistic questions relating to interventions at different stages of IVF might be more suitably answered using stage-specific outcomes of treatment.

1.3 How should we analyse IVF outcomes? Methods and models in use.

If our goal is to predict IVF outcomes for different kinds of patients under different treatment variants, identifying suitable measures of IVF response is only one half of the story. The second challenge is to work out what to do with them. We turn here to the question of how to analyse IVF data, with an informal review of methods employed in the literature. In a clear parallel with our discussion of IVF outcomes, the choice of analytic method will depend on what it is we want to show or find out using data, and who our target audience is. Just as we delineated efficacy and effectiveness in our discussion of outcome measures in section 1.2, we recognise here that different study designs and analysis strategies will be appropriate according to the flavour of our questions. Another distinction that we make is between methods for predicting outcomes of IVF and those for estimating causal effects of patient and treatment characteristics. We will refer to the latter group with the label of 'explanatory' models. We will see that the multistage, repetitious nature of IVF results in a complex multilevel data structure. In the following, it will be necessary to identify where this might represent an obstacle and where it can be leveraged to improve our understanding of IVF. While we will discuss some subtle and challenging methodological features of IVF data, we are equally interested in simple descriptive methods for IVF. The question of how to summarise IVF response data is closely related to the question of the appropriate outcome measure however (once we know which numerator and denominator we are interested in, it is trivial to calculate a proportion), and we have covered many of the relevant considerations in section 1.2. Following the literature, we will use the term 'cycle' to refer to an IVF attempt beginning with the stimulation of the ovaries and ending at treatment failure up to and including the first fresh embryo transfer, or the birth of a child from that embryo transfer. Wherever we discuss frozen transfer cycles, we explicitly refer to these as such.

1.3.1. Models for single treatment attempts

Many researchers present predictive or explanatory models for individual treatment cycles, thereby avoiding (or at least, ignoring) some of the methodological issues arising from the multiplicity of treatment attempts. Baker et al (2010), Templeton et al (1996) and Nelson and Lawlor (2011) all present multivariable logistic regression models for individual cycles, with pregnancy or live birth as the cycle-level outcome. Although their datasets include multiple transfers per patient, these are treated as independent observations. For each of these, frozen transfers are excluded from the analysis.

Although these approaches have the virtue of simplicity, they are not useful for the purposes of predicting or estimating effects on the outcome after a realistic course of treatment. Given that patients are likely to be interested in outcomes over multiple treatment cycles, models based on single cycles of treatment may not be particularly relevant. The exclusion of frozen transfer cycles raises additional doubts about their applicability to treatment programmes as actually realised. Furthermore, wherever participants contribute multiple cycles to the dataset, treating the observations as independent has implications for the validity of inferences made from the models (Vail and Gardener, 2003). To the extent that we are interested in using the models to make point predictions, this may not be particularly worrisome, as we would expect regression coefficients to be generally unaffected. Predictive intervals, which are used to quantify the degree of certainty in the model predictions, may however be too narrow as a result of ignoring the correlation between repeated cycles. Similarly, the confidence intervals we use to make inferences about the model parameters will be incorrect.

Some of the techniques presented as models of treatment cycles would more accurately be described as models for segments of treatment cycles. For example, (Baker, et al., 2010) include intermediate outcomes (such as number of embryos transferred, with a minimum value of 1) as predictors in the model. This restricts the analysis to those patients who actually reached the transfer stage, having had a successful response at the stimulation and fertilisation stages (see also Sunkara et al., (2011), for an example including only cycles where eggs were retrieved but not necessarily fertilised). This might be a reasonable strategy where there is an interest in efficacy, as focusing on part of the process will reduce noise introduced at the other stages. By contrast, Templeton et al., (1996) and Nelson and Lawlor, (2011) included only pre-stimulation variables as predictors, as their purpose was to create prognostic models for patients and clinicians prior to the start of treatment.

1.3.2. Modelling the multiple stages

As we saw earlier, a single cycle of IVF treatment consists of several stages (Figure 1, section 1.1). The treatment is necessarily dynamic, with outcomes of earlier stages determining the range of options available subsequently. Outright failure at an earlier stage (eg: due to poor ovarian response, failure to fertilise oocytes) is critical, resulting in the end of the cycle. There are three common approaches to handling the multistage nature of IVF: using only baseline and start-of-cycle treatment variables as predictors of cycle outcome (avoiding the issue of how to model the multiple stages); presenting conditional models that assume successful outcomes in earlier stages and adjusting for intermediate outcomes in the analysis.

The first of these represents a black box approach, where the dependencies between the different treatment stages are not modelled. Such approaches can be used to predict outcomes of cycles on the basis of baseline and initial treatment variables. If we are interested in questions of efficacy and mechanism however, ignoring the multistage nature of IVF treatment in this way may be disadvantageous for several reasons. Explicit modelling of IVF as a multistage process may elucidate the relationships between treatment stages as well as the differential effects of patient characteristics at different points in the sequence. A multistage model of this sort would allow for the investigation of hypothetical modifications at each stage of the treatment process through simulation. Additionally, an explicit model of the multiple stages might allow for stage-specific modelling of the dropout process. Ultimately, a model incorporating intermediate stages of treatment might allow for better predictions of eventual clinical outcomes. Accordingly, multistage methods may have applications beyond mechanistic investigations. The second approach of the three listed above conditions on success in earlier stages of treatment by leaving out patients who failed early on. As noted above, this restricts the applicability of any estimates and predictions obtained from the model to patients who have reached a certain point in the process. It may be possible however to develop a series of stage-specific and possibly conditional models, and to combine these in a larger framework in order to capture the complete treatment cycle. This approach would require competent models for each stage in the process.

The third approach involves adjusting for intermediate treatment outcomes in the analysis. For example, Cai et al., (2011)used logistic regression to investigate predictors of clinical pregnancy following IVF. They used a bootstrap stepwise variable selection technique to identify a predictive set of variables from an initial set of 27. This set included information about the stimulation protocol, patient characteristics, and intermediate outcomes such as counts of follicles, number of eggs retrieved and fertilised and the number of good quality embryos. However, adjusting for mediating variables may obfuscate the relationship between the outcome and predictors appearing earlier in the causal pathway. For example, both a patient's age and the stimulation protocol used act upon the eventual outcome via the stimulation response. If we wanted to estimate the effect of the stimulation protocol on pregnancy, it would be inappropriate to adjust for variables measuring the stimulation response by including them as covariates in regression models. Techniques from the causal inference literature, such as marginal structural models (Robins, 2000), or structural equation models may be better suited to the task of modelling the complex relationships between patient characteristics, treatment and both intermediate and clinical outcome variables. A sketch of a structural equation modelling representation of the IVF process is presented later in this section.

Several models have however been developed specifically for the purpose of accomodating the multistage nature of IVF. Maity et al., (2014) present two such approaches. The first is a discrete time-to-event model with observations at the level of 'failure opportunities'. They allow for up to three failure opportunities in each cycle; there can be failure of embryo implantation, failure to achieve chemical pregnancy and then there can be spontaneous abortion of the fetus. Consequently, this particular example again assumes a successful stimulation response. Only those patients who did not fail at the previous opportunity feature in the risk set for the next. A patient-specific random scalar is used to account for the nesting of failure opportunities within patients. The authors note however that parameter estimates obtained from the model may be biased due to the fact that the cluster sizes are associated with the outcomes. The second approach presented by the authors is a transition model in which each failure opportunity is modelled as a function of what has happened to the patient previously. This consists of using a fixed effect logistic regression model with failure at the present stage as the binary outcome. Patient history is captured by covariates, including failure type (current stage in the process) and cycle number, so that multiple cycles can be incorporated. One limitation common to both of these methods is that they do not allow for different covariates for each failure type. They do however allow for covariates to have differential effects across stages through the inclusion of interaction terms between the covariates and failure type.

Penman et al., (2007) describe an extended continuation ratio model for an IVF cycle. In this framework, the maximum stage reached by the patient is treated as an ordinal outcome, and the probability of failing at a particular stage given that failure does not occur in the preceding stages is modelled. This approach can therefore be understood as equivalent to a discrete time-to-event model, so that exponentiated coefficients can be interpreted as discrete hazard ratios (Harrell, 2014). As for the techniques presented by Maity and colleagues, the different stages must share covariates, but can incorporate interaction terms between stage and covariate. The five stages included in the model are egg collection, fertilisation, transfer, pregnancy and live birth. This model therefore incorporates the earlier stages of the process and allows for cycle cancellation due to inadequate ovarian response. However, the response to ovarian stimulation is reduced to a binary outcome according to whether or not eggs were collected. This omits prognostic information relating to the number and quality of the eggs retrieved. A superior approach might be to model the response to stimulation as a count of eggs retrieved, or as a prognostic index incorporating information on number and quality of oocytes.

This latter point raises the possibility that models capable not only of incorporating intermediate outcomes, but moreover outcomes of different response types (eg: count, binary and continuous responses) may be needed to adequately capture the multistage nature of IVF. Models for mixed outcomes do not appear to exist in the IVF literature, but have been developed and applied in other research areas. Dunson (2000) described a class of Bayesian latent variable models for clustered mixed outcomes in reproductive toxicity. This approach can accommodate correlations between lower-level units nested within higher-level units in a multilevel data structure, and between multiple mixed outcomes for each lower-level unit. An application is presented where breeding pairs of mice (higher-level units) have up to five litters (lower-level units), and interest is in two related outcomes: whether or not the time to birth is abnormal (a binary outcome) and the number of pups in the litter (a discrete count outcome). The outcomes are related to underlying Poisson variables, with means depending on shared latent variables, inducing dependency. Dunson et al., (2003) would later extend this work to allow for the problem of informative cluster sizes. A general framework for multilevel models containing multivariate mixed responses had previously been outlined by Goldstein (2003) and was later developed by Goldstein et al., (2009). These methods raise the possibility of jointly modelling the various intermediate and final outcomes of IVF cycles across multiple levels of the data structure while accounting for the fact that the number of lower-level units (which may be eggs, stages or cycles) within a cluster (stimulation episodes, cycles, patients) is related to those outcomes.

Other methods for modelling multistage processes exist, such as methods for the estimation of dynamic treatment regimes (Chakraborty and Murphy, 2014). While these techniques appear superficially applicable to multistage IVF treatment, a brief review of the literature in this area suggests that they are best suited to short sequences of binary treatment decisions and responses. We do not consider them further here.

1.3.3. Structural equation modelling approaches

Limitations of the models described above include the inability to accommodate or investigate the complex dependency structure between patient characteristics, different stages of the treatment process and the multiple mixed outcomes occurring during the process. A structural equation modelling (SEM) approach allows for complex relationships between covariates and outcomes, including the possibility that covariates may act upon outcomes indirectly (Bielby, et al., 1977, Fox, 2006). Furthermore, outcomes may follow different distributions and depend upon each other (Rabe-Hesketh, et al., 2004). This framework might be sufficiently flexible to accommodate the complexities of IVF. We present a sketch of what this might look like here. It should be stressed that the example presented here is intended to be expository (and exploratory), in order to show how SEM methodology might be used to capture the structural complexity arising from the multiple stages constituting a single cycle.

A SEM contains ‘exogenous’ and ‘endogenous’ variables (Fox, 2006). The former appear only as explanatory variables in the model. Each endogenous variable appears as the response variable of a structural equation, but may appear as an explanatory variable in others. Accordingly, traditional outcome variables may act as predictors of other outcomes. Variables in the model may be observed or latent, representing unobserved or postulated constructs (Rabe-Hesketh, et al., 2004).

Figure 3 shows a simple example of an SEM for an IVF cycle, represented by a path diagram (Greenland, et al., 1999, Pearl, 1995). The path diagram contains the exogenous variables age and stimulation protocol (although in reality, stimulation protocol might be partially selected on the basis of age), and the endogenous variables stimulation response, pregnancy and live birth. Bidirectional arrows represent covariances, which are not given causal interpretations. Unidirectional arrows represent direct effects of one variable (at the origin) on another (at the head). In this example, a patient's age and the stimulation protocol have direct effects upon the stimulation response, representing the fact that a patient's ovarian reserve may deplete with age and that the specific regimen used may influence the number of eggs retrieved. Age and stimulation protocol affect pregnancy indirectly via the response to stimulation. They also act directly on pregnancy, representing a possible diminishing uterine receptivity with age and the fact that ovarian stimulation may render the uterus inhospitable to transferred embryos. Similarly, age is allowed to act upon live birth outcome indirectly via the intermediate outcomes stimulation response and pregnancy, and also directly, representing the fact that a patient's capacity to carry a pregnancy to birth may diminish as they get older. By contrast, stimulation protocol does not affect live birth rate in the model, other than through indirect effects via the intermediate outcomes. Additionally, stimulation response, which incorporates the quality of oocytes retrieved, is allowed to act directly upon live birth, representing the possibility that better quality oocytes may produce superior embryos that are more likely to be carried to term. A single latent error term exists for each endogenous (response) variable, representing the effects of unmeasured variable and measurement error. Covariance between the errors for the endogenous variables is posited.

An SEM can then be represented as a set of simultaneous equations, and model coefficients can be estimated using a number of approaches. As noted above, this presentation is intentionally simplistic and some of the details have been left deliberately vague. Notably, the forms of the predictor and response variables and the distributions of corresponding error terms have not been specified. A particular challenge may arise from the fact that censoring occurs due to failure at earlier outcomes. However, the framework offers scope for considerable expansion and improvement. Count and binary outcomes can be incorporated in a generalised SEM framework (Rabe-Hesketh, et al., 2004). Additional exogenous and endogenous variables can be incorporated to more fully describe the treatment process and the patient characteristics influencing the outcome at each stage. Models can also be specified in a multilevel framework, in order to accommodate clustering of repeated observations. Given that the emphasis in this framework lies in estimating the relationships between variables, we anticipate that this approach would be more suitable for answering causal questions relating to efficacy and mechanism rather than those related to prediction. Although these comments are necessarily tentative at this stage, an SEM approach to modelling the multistage IVF cycle would appear to warrant further attention.

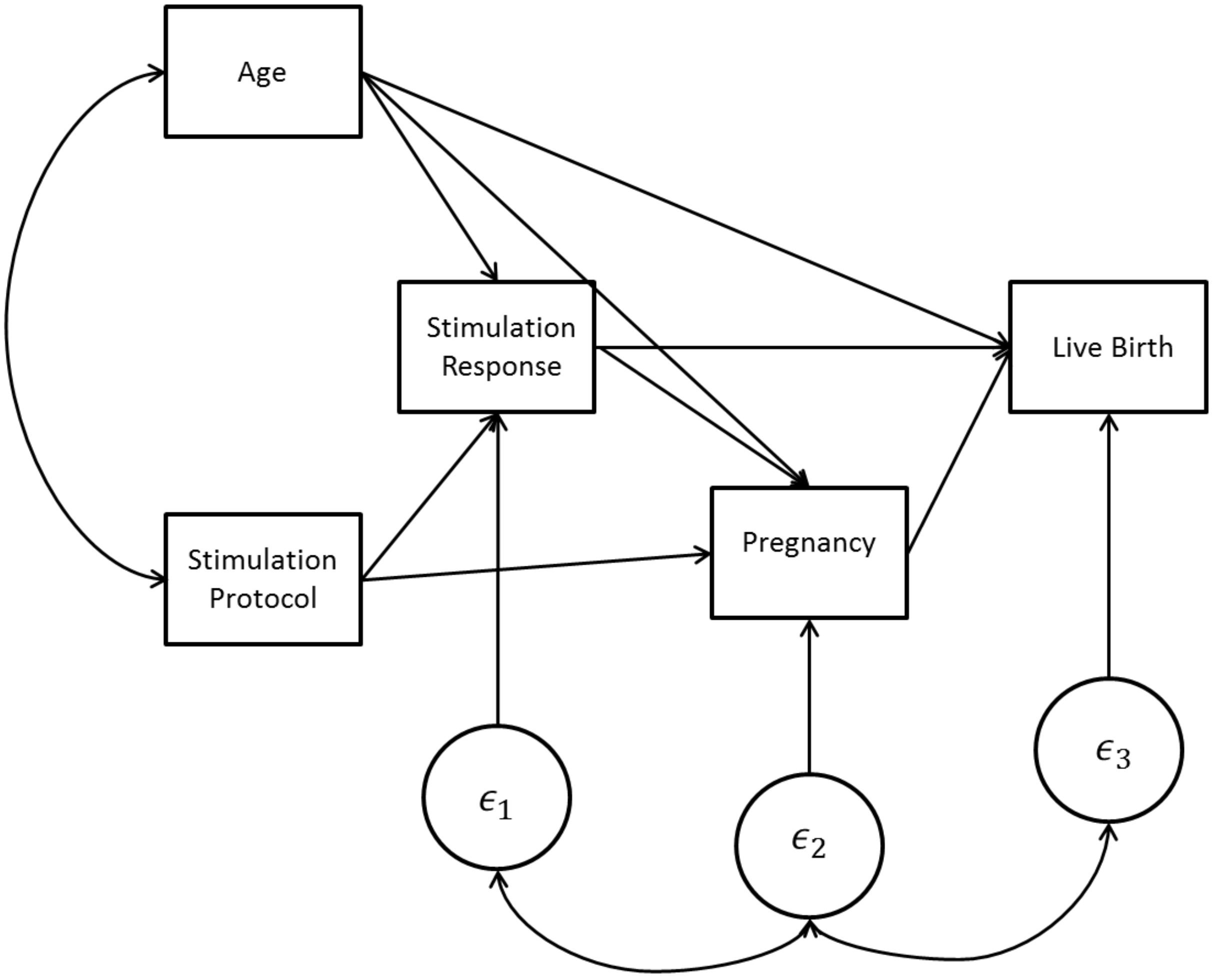


Figure 3: Sketch of a possible structural equation modelling approach for an IVF cycle. Rectangles represent observed variables, while circles represent latent variables. Unidirectional arrows denote causal relationships with the affected variable at the head. Bidirectional arrows denote covariances, which are not given a causal interpretation.

1.3.4. Models for the stimulation stage of IVF

Several of the aforementioned approaches to modelling the IVF treatment cycle as a multistage process require the explicit modelling of the outcomes at each stage. For example, the simple SEM example described in the preceding section would require adequate models for each of the stimulation response, the odds of achieving pregnancy and the odds of achieving a live birth, in terms of patient and treatment characteristics and the preceding outcomes. The second and third of these may be modelled quite straightforwardly using standard models for binary outcomes (although complexity arises once we distinguish multiple births, see 1.2.5). The matter of how to model the stimulation response is potentially more challenging, as both the number and quality of retrieved oocytes may be relevant as predictors of success in later stages. Treating the stimulation response as a binary success/ failure outcome (Penman, et al., 2007) or omitting this stage from consideration entirely (Maity, et al., 2014) omits prognostic information and is likely to reduce the relevance and predictive validity of the model. Furthermore, the stimulation stage appears to be particularly important, as the response determines not only the outcome of the fresh cycle but also the number and outcome of subsequent frozen transfers. In this section, we will review the models that have been presented for the ovarian response to stimulation.

The ovarian response is sometimes categorised on the basis of the number of eggs retrieved. Mohiyiddeen et al., (2013b) categorised the response as poor (fewer than 4 eggs), normal (4 to 20 eggs) or overresponse (more than 20 eggs). They then fitted two logistic regression models looking at predictors of each of over and poor response compared to normal response. The usual criticisms of categorisation are applicable here; information contained in the original variable is lost and the arbitrary division into discrete categories is unlikely to provide a good representation of the underlying reality (Altman and Royston, 2006). La Marca and Sunkara, (2014) presented a review of markers of ovarian reserve for predicting stimulation response. They restricted their review to studies reporting cut-offs for the markers, so it is unsurprising that the included studies categorised ovarian response; this is a necessary step in calculating the predictive accuracy of a marker using a method such as a ROC curve (although again this approach wastes information). They noted considerable variation in the cut-off used to define poor response, highlighting the lack of consensus around this topic. Number of eggs retrieved is not always categorised; for example, (Gaskins, et al., 2014) modelled this as a count variable using Poisson regression. Where the raw egg count is used as the measure of stimulation response however, a common statistical error in IVF trials is to exclude participants who had their stimulation cancelled due to an anticipated poor response (eg: Arce, et al., 2014, Cavagna, et al., 2006, Jayaprakasan, et al., 2010, Nyboe Andersen, et al., 2017). This is fatal to the estimation of a treatment effect, both because it undermines the balance produced by randomisation and because it omits those participants who were on course to have a low number of eggs. An instructive example of this error is provided in a trial of a personalised dosing algorithm versus a standard dose of drug for ovarian stimulation (Nyboe Andersen, et al., 2017). The authors claimed an advantage of the algorithm in achieving an optimal egg yield, even though the rate of anticipated poor responders (who they excluded from the calculation) was higher in this arm of the trial. This illustrates the fact that censored egg counts must be considered when modelling the stimulation response, and more generally the scope for methodological error introduced by inappropriate denominator selection.

In 2011, ESHRE presented the Bologna criteria for defining poor response to stimulation (Ferraretti and Gianaroli, 2014). Their actual focus was more on defining a 'poor responder' rather than what should be considered a poor response to a particular stimulation episode. They describe the stimulation as a test of the ovarian reserve (the dormant eggs which are the target of the procedure) noting that an ideal test would capture both the size of the primordial follicle pool and the reproductive competence of the oocytes. This perspective emphasises the point that both the number and quality of oocytes harvested following the stimulation procedure are relevant components of the outcome. A good model of stimulation response might capture both of these aspects, prompting the question of how this could be achieved. An argument could be made that number of oocytes retrieved might adequately capture egg quality, as larger numbers of eggs might be expected to result in larger numbers of embryos, the best of which can be selected for transfer. However, the relationship between number and quality of oocytes retrieved is unclear, so that it may be possible that large numbers of low quality oocytes could be obtained (or even that the production of greater numbers of oocytes in response to stimulation has detrimental consequences for egg quality). Another complication is the fact that the desirable number of oocytes will vary according to the characteristics of the patient and the stimulation regime used; a low number of oocytes might be considered a success following a mild stimulation protocol (Ferraretti and Gianaroli, 2014). As a result, it might not be possible to standardise the measurement of optimal egg yields across patient populations.

If quality is to be explicitly modelled in addition to quantity of eggs, a further question is how this can be measured. Systems based on morphology exist, but remain controversial and are only applicable when the eggs are stripped and fertilised by injection with sperm (intracytoplasmic sperm injection, or ICSI, Mohiyiddeen, et al., 2013a). One possibility is to use maturity of the oocytes for this purpose: Guerif, et al., (2009) reported the proportion of metaphase-II (mature) oocytes in the yield; Mohiyiddeen, et al., (2013a) used the metaphase-II oocyte output rate (MOR), defined as the ratio of metaphase-II oocytes to the pre-stimulation antral follicle count, where antral follicles house the eggs within the ovaries.

One possible approach to modelling stimulation response is to combine information about number and quality of oocytes into a single prognostic index, obtained from a multivariable model including measures of ovarian response as predictors of clinical outcomes (pregnancy, live birth). Several studies have looked at the association between clinical outcomes and stimulation response. One study looked at characteristics of the stimulation procedure (the total gonadotrophin dose per cycle, the number of oocytes retrieved and the gonadotrophin dose per oocyte) as predictors of pregnancy and implantation rate (Kailasam, et al., 2004). Other, similar studies have related number of eggs to live birth (Sunkara, et al., 2011), and the ratio of preovulatory to antral follicle count to clinical pregnancy (Gallot, et al., 2012). Multivariable predictive models of this sort would be required to obtain regression coefficients which could be used to calculate the prognostic index. Combining a multivariate response into a single index is likely to have detrimental consequences for efficiency, however.

Another possible approach would be to model the response to stimulation as a multivariate outcome, with components describing each of quality and quantity. Several of the techniques discussed above can accommodate such multiple, mixed outcome variables. In particular, the method of Dunson, et al., (2003) can be used to simultaneously model the number of eggs retrieved with a measure of the quality of each egg. This is similar to an example given by Dunson where the overall size of a litter of mice is modelled jointly with the birthweight of each fetus.

Whichever solution turns out to be appropriate, it seems clear that the development of a model capturing the information provided by this crucial stage of treatment is a necessary precursor to modelling the IVF cycle as a multistage process.

1.3.5. Extending to multiple treatment cycles

The methods described above are applicable to single cycles of treatment, possibly broken into component stages. Since patients typically undergo multiple cycles of treatment, there is a need for methods that can handle this repetition.

The fact that repeated cycles undertaken by a patient can be expected to be correlated has been widely recognised (Missmer, et al., 2011, Penman, et al., 2007, Roberts and Stylianou, 2012, Hogan and Blazar, 2000, Maity, et al., 2014, Vail and Gardener, 2003, Gaskins, et al., 2014, Hirst, et al., 2011, Jonsdottir, et al., 2011). Failure to account for this correlation may result in erroneous inferences and predictions. If multiple treatment cycles are being modelled, it is therefore necessary to employ techniques that can handle the clustering of repeated observations on a single patient.

One approach to modelling the correlation between cycles is to include a patient-specific random effect in the model, representing unexplained heterogeneity between patients. This multilevel modelling approach has been used to extend logistic regression models (Gaskins, et al., 2014, Hirst, et al., 2011, Hogan and Blazar, 2000, Missmer, et al., 2011), a multistage continuation ratio model (Penman, et al., 2007), a multistage discrete time survival model (Maity, et al., 2014) and embryo-uterine models (Roberts and Stylianou, 2012) to multiple treatment cycles. An alternative approach is to use a method based on generalised estimating equations (GEE), which allows for the correlation structure between repeated measurements to be specified. Logistic GEE models with compound symmetric correlation structure have been used to model repeated IVF treatments (Jonsdottir, et al., 2011, Missmer, et al., 2011). Although there are subtle differences in the interpretations of parameters obtained from conditional random effects models compared to marginal GEE approaches, both represent valid approaches for the analysis of multiple-cycle IVF data.

Methods using a patient-specific random effect or a compound symmetric correlation structure tacitly assume that the correlation between multiple cycles on a single patient remains constant, or equivalently that unexplained differences between patients are attributable to time-invariant unmeasured characteristics. This is described by Penman et al., (2007) as an `infertility index'. This assumption of constant correlation between treatment cycles regardless of temporal proximity is questionable. More realistic models may incorporate techniques from the longitudinal data analysis and growth curve modelling literatures, including the inclusion of random slopes to allow for a patient's likelihood of success to follow an individual trajectory over time and the specification of more nuanced correlation structures; for example an autoregressive structure allowing correlation between cycles to decrease as a function of increasing time (Diggle, et al., 1994). It is unclear exactly why these techniques have not been employed in the analysis of IVF data. One reason may be that IVF data hasn't been recognised or conceptualised as longitudinal data, although as an explanation this begs the question. Another explanation might be the complexity of implementing time-varying correlation structures with binary endpoints. And a third might relate to unavailability of repeated-cycle data. Whatever the reason, little has been done to incorporate ordering effects into models for repeated cycles, beyond the inclusion of a covariate representing the number of previous attempts (Nelson and Lawlor, 2011, Roberts and Stylianou, 2012, Templeton, et al., 1996). The fact that number of attempts appears to be predictive of cycle outcome even after adjusting for other prognostic characteristics might suggest that there are unmeasured time-varying factors or selection effects that contribute to the chance of success. A 2010 review article considered the supposition that repeated ovarian stimulation diminished ovarian reserve (Luk and Arici, 2010). The authors concluded that there was limited evidence on this topic, but that there did not appear to be declines in ovarian response over three cycles of treatment. Beyond three cycles of treatment, the authors noted difficulties separating effects of repeated stimulation from effects of increasing age. Elsewhere, it has been argued that the decrease in odds of success with increasing cycle number appears to be small (Roberts and Stylianou, 2012) and that apparent reduction in success with increasing cycle number may be attributable to a selection effect resulting from the fact that patients with the worst prognosis are those who require the largest number of attempts (Dias, et al., 2008). Longitudinal models of IVF treatment may provide a framework to investigate these points further.

Prognostic characteristics may vary over time. This presents a particular problem where interest lies in using multiple cycle data to estimate the effect of a treatment or exposure. If exposure probability depends on these characterisics, then time-varying confounding may occur. Hogan & Scharfstein, (2006) presented a model to estimate the causal effect of hydrosalpinx on embryo implantation in the presence of time-varying confounding. Their method is based on inverse weighting of observations by propensity scores. A further complication arises from the fact that adequate modelling of frozen cycles may require the inclusion of different covariates compared to fresh cycles. This matter has not been addressed to date; existing methods include use of common covariates for fresh and frozen transfers, or excluding frozen transfers entirely.

An interesting recent approach to prediction of the cumulative IVF outcome after multiple treatment attempts was provided by McLernon and colleagues (2016a). They presented two models. In the first, they used only pre-treatment variables as covariates, to provide a prediction to prospective patients deciding whether to commence treatment. In the second, they used information about the responses in the first treatment attempt to predict outcomes in subsequent attempts. This second model can be used to counsel patients considering whether to undergo additional treatment. As discussed in section 1.2.4 the cumulative outcome of treatment over multiple attempts is probably the most relevant measure for patients, making it an appropriate choice of endpoint in a prediction model freely available online (McLernon, et al., 2016b).

1.3.6. Partial observability of transferred embryos

When multiple embryos are transferred simultaneously and only some of these implant, it is unknown which of the embryos were successful. When covariates are measured at the level of embryos, an additional challenge stems from the fact that the outcomes of interest occur at the patient, rather than at the observation, level (Roberts, 2007). Embryo-uterine (EU) models were developed in order to address this issue of partial observability and to allow for the inclusion of embryo-level covariates (Zhou and

Weinberg, 1998). Briefly, EU models assume that for an embryo to implant, the embryo must be viable and the uterus must be receptive. Covariates are entered into one of two logistic sub-models, with viability of the embryo and receptivity of the uterus as binary responses. Maximum likelihood (Roberts, 2007, Roberts and Stylianou, 2012) and Bayesian (Corani, et al., 2013, Dukic and Hogan, 2002) approaches to fitting these models have been developed. In any situation where embryo-level covariates are of interest and multiple embryos are transferred simultaneously, EU models represent the preferred method of analysis. It is possible to use a Logistic-Normal mixed effects model (Roberts, 2007), although interpretation of coefficients is complicated by the fact that cycle outcomes do not occur at the embryo level. It is possible that concern over multiple births will lead to the adoption of elective single embryo transfer as the default treatment strategy, which would render this a moot point. At present however, transfer of multiple embryos occurs in the majority of cycles in the UK, so that this issue is likely to remain relevant for the foreseeable future (Human Fertilisation and Embryology Authority, 2016).

1.3.7. Dealing with drop out

Many couples discontinue treatment before achieving a successful result. Due to the multistage and multi-cycle nature of IVF, discontinuation may occur midway through a cycle or following the completion of a cycle that resulted in failure. It has been noted, usually in relation to the calculation of cumulative live birth or pregnancy rates, that this drop out presents a challenge when analysing IVF datasets. The estimation of success rates using the Kaplan-Meier method produces biased estimates, as an underlying assumption of this approach is that censoring is independent of outcome. A considerable literature exploring the reasons for drop out in IVF has emerged, and active censoring of the patient due to poor prognosis is commonly noted to be prevalent (Brandes, et al., 2009, Olivius, et al., 2004, Verberg, et al., 2008). Studies investigating predictors of drop-out have identified factors which are also prognostic for treatment outcome, such as age (Luke, et al., 2013, Troude, et al., 2014) and stimulation response (Marcus, et al., 2011, Pelinck, et al., 2007, Troude, et al., 2014). The supposition that those patients who abandon treatment would, if they did choose to proceed, have the same chance of success as those who do continue therefore appears to be implausible.

Recognising this point, several researchers have calculated rates based on alternative assumptions. A 'pessimistic' approach assumes that patients who drop out would have a zero probability of success if they continued treatment (Stewart, et al., 2011a, Malizia, et al., 2009, Verhagen, et al., 2008). This is clearly overly conservative, and some researchers have adopted the approach of presenting this conservative rate together with the rate based on the assumption of independent censoring, reasoning that the true rate will lie somewhere between these two bounds (Lintsen, et al., 2007, Malizia, et al., 2009, Stewart, et al., 2011a). Another approach is to assume that those who were censored for medical reasons had zero chance of success, while those who were censored for other reasons were as likely to succeed as those who continued (Verhagen, et al., 2008). This latter approach therefore makes separate unrealistic assumptions for those who were actively and passively (voluntarily) censored, and is not obviously an improvement on the methods hitherto described.

Several attempts have been made to accommodate censoring that occurs in IVF. Soullier, et al., (2008) used multiple imputation (MI) to estimate cumulative live birth rates in the presence of drop out. An assumption of this approach is that missing outcome data are missing-at-random (MAR) (Rubin, 1976). This means that differences between missing and observed outcomes can be explained using observed data (Sterne, et al., 2009). The authors included the variables IVF unit, age at aspiration, number of oocytes retrieved, and total number of embryos (transferred plus frozen) as predictors in the imputation model, on the grounds that these are frequently cited as predictors of outcome in the literature. It is unclear that a model with so few predictors would be sufficient to provide reasonable predictions of the missing values. The literature pertaining to predictors of IVF outcome and reasons for drop out might enable superior missing data models to be constructed. This might highlight relevant variables to include and whether or not these are likely to be sufficient to account for systematic differences between patients who do and do not abandon treatment. For example, if a common reason for drop out is financial expense (which was cited as a reason by 46% of respondents to an internet survey of 80 users of an independent infertility website, Marcus, et al., 2011), then differences may exist between those who abandon treatment and those who do not in terms of socio-economic status. This might prompt us to consider whether or not such differences could be accounted for by the available variables. We should be wary of unvalidated prediction models appearing in the literature however, as reported associations may be spurious or otherwise non-generalisable. For example, a systematic review of 22 studies of reasons for and predictors of drop out in fertility treatment found that no predictors were consistently associated with drop-out (Gameiro, et al., 2012).

Another attempt to address the drop-out over multiple treatment cycles was made by Hogan and Scharfstein, (2006), who noted that attrition depends heavily on prior outcome. They presented a causal model with per-embryo implantation rate as the outcome, based on a propensity weighting method. This approach again assumes that outcome data are MAR, so that missingness is fully explained by observed data. As for the preceding MI example, this approach therefore requires a good understanding of the reasons for drop out and the availability of variables needed to produce a satisfactory missing data model.

Concerns over informative censoring also arise when considering models for the multiple treatment stages within a cycle. Patients may abandon treatment mid-cycle. Poor response at an earlier stage may preclude completion of the treatment cycle. At present, little attention has been paid to the issue of missing data for multistage prediction models, with most researchers either ignoring or acknowledging and dismissing the problem. For example, when discussing the continuation ratio model for IVF, Penman et al., (2007) made the (weak) argument that patients drop out for a variety of reasons, and that missingness could therefore be considered ignorable. Maity, et al., (2014) acknowledged that estimates from their discrete time-to-event model for the multiple stages of IVF could be biased due to the fact that the number of observations made on a given patient was itself informative.

Methods for dealing with informative censoring have been developed in other topic areas, and are likely to be applicable or modifiable for the case of the multistage IVF cycle. Dunson, et al., (2003) presented a method for the case where the number of observations is informative, involving the joint modelling of several mixed outcomes together with the cluster size. This was extended by Goldstein, et al., (2009) to provide a general class of models capable of handling missing data in multilevel contexts with multiple mixed outcome types. This approach might allow for missingness to be modelled at the level of stages within a multiple-cycle framework. A model for longitudinal binary response data with informative missingness that makes use of the continuation ratio model to represent the drop out process has been described by Ten Have, et al., (1998). A method for relaxing the independent censoring assumption in the Cox proportional hazards model has been presented by Jackson, et al., (2014). These latter two examples may be useful for modelling the drop-out process over full or extended courses of treatment.

1.3.8. Time-to-event models

As has been discussed in section 1.2.7, the distressing nature of subfertility and IVF treatment means that the expected time to achieve success is likely to be an important piece of information for patients. This motivates the use of survival models for time-to-event outcomes in IVF. As noted in the previous section, Kaplan-Meier methods are commonly used to estimate cumulative live birth or pregnancy rates over several cycles of treatment. Typically, cycle started or completed is used as the unit of time in these analyses. Similarly, Missmer, et al., (2011) use cycle completed as the unit of time in a discrete time survival model, which they note to be equivalent to performing unconditional logistic regression with cycle number as a covariate. In reality the time taken to undergo a set number of cycles may vary drastically between patients, so that results presented using cycle number as the unit of time may be difficult to interpret. Moreover, due to the lack of a common timescale in these approaches, comparisons between treatment programmes are not meaningful (Daya, 2005). These problems are compounded in discrete time-to-event models for handling the multistage nature of IVF, where treatment stages are the unit of time[1] (Maity, et al., 2014, Penman, et al., 2007). Models based on real time are therefore likely to be more relevant to patients (Daya, 2005).

Although not as common as analyses based on discrete-time approaches, analyses using real-time survival methods do appear in the IVF literature. For example, Verhagen, et al. (2008) presented cumulative pregnancy rates based on real-time over a 9 month period. Lintsen, et al., (2007) used the Cox proportional hazards model with time from commencement of treatment to ongoing pregnancy as the outcome of interest. They

[1] It should be noted however, that these models were not intended for the purpose of forecasting expected times to a successful result.

included baseline covariates and appear to have treated drop outs as failures. This highlights the need for time-to-event methods capable of incorporating more realistic drop-out assumptions and possibly time-varying covariates. Another possible area for development lies in methods for the joint modelling of repeated measures data and event times (eg: Hogan and Laird, 1997). These can be used to estimate the effect of a longitudinal measure on a time to event outcome while allowing for unmeasured confounding between the two (Rizopoulos, 2012).

1.3.9. Summary of section 1.3

Attempts have been made to develop models capable of accommodating the multistage and multicycle structure of IVF treatment. Methods for multiple cycles generally assume that correlation between cycles is constant, or that unexplained heterogeneity is attributable to time-invariant unmeasured factors. Methods from the longitudinal data literature may offer ways to relax this assumption, although it is worth questioning whether or not this is likely to produce substantively different clinical conclusions. If not, such an exercise would have little practical benefit. Given the finding of Roberts and Stylianou, (2012), that correlation due to unmeasured factors across repeated cycles of IVF was modest, the value of pursuing this matter further is somewhat doubtful.

The matter of how to model the multiple stages of an IVF cycle has received relatively little attention. The methods that have been proposed have limitations, most notably in the treatment of the critical stimulation stage. Adequate modelling of the stimulation response would seem to represent an important first step in modelling the full treatment cycle. Several strategies for modelling the ovarian response have been outlined in brief. Once adequate stage-specific models have been established, the question of how to incorporate these into a model for a complete (fresh) cycle requires consideration. Although it has received relatively little attention to date, the idea of developing multistage models including the various stages of IVF treatment is interesting, particularly in relation to questions of efficacy. Existing methods dichotomise the stage-specific responses, representing a waste of information. Methodological areas have been identified which may be useful for this purpose, including joint modelling and SEM approaches.

Although discrete time to event methods appear in the IVF literature, these have the potential to be misleading due to variation between patients and between centres in the timings involved. Survival models based on real-time are occasionally used, but more consideration is needed in relation to time-varying covariates and unrealistic censoring assumptions.

Appropriate methods for handling informative drop-out have been discussed for multiple cycle models, but these are rarely implemented. Where methods are used, for example in the calculation of cumulative rates, the underlying assumptions are dubious. The matter of modelling the drop-out process in the multistage context has yet to be considered in the literature. Methods for the joint modelling of cluster size and outcomes derived in other topic areas may be modifiable for the purpose of describing attrition within a cycle.

1.4 Conclusions of the literature review

The objectives of the present literature review were to establish and summarise areas of consensus and dissent on the topic of outcome measurement in IVF, and to critically review analysis strategies appearing in the literature. On the basis of this informal review, we believe that the search for a single measure of success is a castle in the air. The appropriate choice of measure will depend on the purpose and intended audience, and in fact it might be necessary to use a suite of several measures to capture different aspects of treatment (effectiveness and safety, for example). The appropriate approach to analysis is also likely to be context-dependent and closely related to the appropriate choice of outcome measure(s).

Wherever a measure is intended to inform current or prospective patients about the likelihood of success, the literature generally supported the view that live birth rates over full courses of treatment are the most relevant outcomes to patients. We believe that this is likely to be true for effectiveness trials in general, although the practicalities of following patients over several cycles might preclude this. Despite its attractiveness as a pragmatic measure, the review highlighted substantial logistical and conceptual difficulties with the use of cumulative birth rate as an outcome. In the absence of long-term datasets linking patients between multiple centres, a more realistic ambition would be to develop models over some prespecified period of time based on single-centre data.

While it might be possible to develop methods that could incorporate multicentre treatment courses, the application would probably have to remain hypothetical. Where cumulative live birth rates cannot be obtained, recording live birth after a single cycle probably takes second place prize (although this will not be suitable for the evaluation of policies based on the number of attempts to offer, eg: McLernon, et al., 2010). If live birth rates cannot be obtained, the ongoing pregnancy rate (for example, after 12 weeks of gestation) is a reasonable surrogate measure (Braakhekke, et al., 2014), provided that the (small numbers) of patients who are observed to have a miscarriage later in the gestational period are not then reclassified as having failed treatment.

Given that methods for multiple treatment cycles already exist (section 1.3.5.) this line of research would only be worth pursuing if those existing models had substantial limitations. Simplistic assumptions relating to correlation structure and missing data mechanisms have been identified as possible weaknesses in analyses generally conducted. However, it is unclear that either of these represent enough of a problem to justify a targeted program of research. The first might not matter in practice (Roberts and Stylianou, 2012) and the second appears to be a matter of raising awareness of existing missing data methods (Soullier, et al., 2008, Sterne, et al., 2009, Ten Have, et al., 1998). An alternative way to analyse multicycle data might be to use time-to-event methods, using realistic censoring assumptions. This might capture the burden of treatment, by rewarding interventions that shorten the route to birth. One concern with survival analyses however is the risk of rewarding aggressive and unsafe treatment strategies, which may result in shorter times to pregnancy and birth. Adverse outcomes such as hyperstimulation syndrome should always accompany time to event effectiveness measures. Given that adverse event rates are generally relatively low in comparison to birth rates however, there is a risk that harmful effects of treatments will go unrecognised due to low power.

While questions of IVF effectiveness and clinical outcome prediction are most vital to patients, most of the challenges we encounter in practice can be tackled using existing statistical methods for answering pragmatic research questions, handling repeated measures and reducing bias due to missing data. The question of how to investigate the internal structure of an IVF cycle is more enticing. Using multistage models might elucidate the effects of treatment decisions and patient characteristics on responses at different stages of the process and possibly result in superior predictions of birth outcomes. Several proposals have been made, although these have severe limitations, including inadequate handling of the stimulation stage, reductionist dichotomisation of response variables, and inability to accommodate informative censoring. In practice, the simplistic treatment of the stimulation stage is probably a symptom of the complexity of representing the ovarian response, as well as the difficulty of combining mixed response types in a single model. Both the matter of how to model the stimulation stage, and how to subsequently combine the sequence of stages into a coherent model for the IVF cycle present substantial challenges, although methods for handling complex relationships between covariates and multiple mixed outcomes have been described. An aspiration for this line of work would be to produce something that could be extended or embedded into a model for multiple treatment cycles.

This literature review has limitations. The relevant body of literature was both disparate and abundant, so that a finite population of studies and research papers could not be identified. The implication of this is that the review cannot be conceived of as ‘complete’ in any sense: the identification and assimilation of relevant research will be an ongoing process which will continue to inform the project strategy. However, the primary purpose of this review was to develop an understanding of the work that has been done to date and to generate ideas for subsequent development. Several avenues of investigation have been opened, and several areas of literature that merit further consideration have been identified. In light of this, a multi-faceted strategy based on the research ideas presented here suggests itself as a suitable programme to adopt from this point.

1.5 Next steps and outline of the thesis structure

Our intention for the remainder of the thesis is to develop statistical methods for the purpose of answering mechanistic questions relating to the internal structure of the multistage IVF cycle. While we have seen some literature discussing which outcome measures should be used to evaluate IVF, the bulk of this literature was concerned with patient-facing outcome measures, such as national performance indicators and endpoints for effectiveness research. The literature on measurement of stage-specific responses for answering mechanistic research questions has not been so well-represented in this review, which could be a product either of sparsity of work in this area or a deficiency in our search, which was conducted in an informal manner. Another way to establish stakeholders' attitudes towards IVF outcome measures would be to see what they are actually using. To this end, we will conduct reviews of the outcomes measures actually in use in different contexts. We anticipate this exercise will have several benefits. First, it will allow us to systematically establish the outcome measures in use for different purposes, which will inform the response variables to include in our models. Second, given the complex structure of IVF, we anticipate that there will be different ways in which we could divide the cycle into discrete stages. A review of the numerators and denominators in use will give some insight into common and natural strategies to carry out this partitioning. Finally, we anticipate that there will be scope to comment on current reporting practices by publishing our reviews. Accordingly, this preparatory exercise might have some direct benefit.

We will require good statistical models for the different stages of the IVF cycle (with the set of stages still to be identified from the review of outcome measures just described). We begin by focussing on modelling strategies for the stimulation stage, which the literature review has identified as a weakness in existing multistage models. We explore this matter by way of an application to real data. Subsequently, and as we identify a suitable set of stages to be modelled, we will investigate how to model and then combine stage-specific responses in order to answer mechanistic research questions.

This program of work comprises two coalescing strands, and this is reflected in the structure of the thesis. In Part II, we detail our research into outcome measurement in IVF, while in Part III we turn our attention to the matter of developing multistage models for IVF. Each of these parts is divided into methods and results sections. Our results are presented in journal article format, and so in our methods sections we present and critically motivate our methodology in greater detail than is permitted given publishing restrictions.

1.6 References for Chapter 1.

Abdalla HI, Bhattacharya S, Khalaf Y. Is meaningful reporting of national IVF outcome data possible? *Hum Reprod* 2010;25: 9-13.

Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332: 1080.

Arce JC, Andersen AN, Fernandez-Sanchez M, Visnova H, Bosch E, Garcia-Velasco JA, Barri P, De Sutter P, Klein BM, Fauser BCJM. Ovarian response to recombinant human follicle-stimulating hormone: a randomized, antimullerian hormone-stratified, dose-response trial in women undergoing in vitro fertilization/intracytoplasmic sperm injection. *Fertil Steril* 2014;102: 1633-U1456.

Baker VL, Luke B, Brown MB, Alvero R, Frattarelli JL, Usadi R, Grainger DA, Armstrong AY. Multivariate analysis of factors affecting probability of pregnancy and live birth with in vitro fertilization: an analysis of the Society for Assisted Reproductive Technology Clinic Outcomes Reporting System. *Fertil Steril* 2010;94: 1410-1416.

Bielby WT, Hauser RM, Featherman DL. Response Errors of Black and Non-Black Males in Models of Intergenerational Transmission of Socioeconomic-Status. *Am J Sociol* 1977;82: 1242-1288.

Bird SM, David C, Farewell VT, Harvey G, Tim H, Peter C. Performance indicators: good, bad, and ugly. *J R Stat Soc Ser A Stat Soc* 2005;168: 1-27.

Braakhekke M, Kamphuis EI, Dancet EA, Mol F, van der Veen F, Mol BW. Ongoing pregnancy qualifies best as the primary outcome measure of choice in trials in reproductive medicine: an opinion paper. *Fertil Steril* 2014;101: 1203-1204.

Brandes M, van der Steen JOM, Bokdam SB, Hamilton CJCM, de Bruin JP, Nelen WLDM, Kremer JAM. When and why do subfertile couples discontinue their fertility care? A longitudinal cohort study in a secondary care subfertility population. *Hum Reprod* 2009;24: 3127-3135.

Cai QF, Wan F, Huang R, Zhang HW. Factors predicting the cumulative outcome of IVF/ICSI treatment: a multivariable analysis of 2450 patients. *Hum Reprod* 2011;26: 2532-2540.

Cavagna M, Freitas GC, Soares JB, Sales AL, Andrade PC, Dzik A. Comparison of 225 IU and 300 IU follitropin-A in a fixed-dose regimen for controlled ovarian stimulation in women aged 35 years or older. *Fertil Steril* 2006;86: S408-S408.

Chakraborty B, Murphy SA. Dynamic Treatment Regimes. *Annu Rev Stat Appl 1* 2014;1: 447-U1025.

Chetkowski RJ. Consumer-friendly reporting of in vitro fertilization outcomes. *Fertil Steril* 2014;101: e7.

Coetsier T, Devroey P, Dhont M, Edwards RG, Evers H, Hagglund L, Handyside A, Gerris J, Koudstaal J, Vilska S *et al.* Prevention of twin pregnancies after IVF/ICSI by single embryo transfer. *Hum Reprod* 2001;16: 790-800.

Corani G, Magli C, Giusti A, Gianaroli L, Garnbardella LM. A Bayesian network model for predicting pregnancy after in vitro fertilization. *Comput Biol Med* 2013;43: 1783-1792.

Davies MJ, Wang JX, Norman RJ. What is the most relevant standard of success in assisted reproduction? Assessing the BESST index for reproduction treatment. *Hum Reprod* 2004;19: 1049-1051.

Daya S. Life table (survival) analysis to generate cumulative pregnancy rates in assisted reproduction: are we overestimating our success rates? *Hum Reprod* 2005;20: 1135-1143.

Dias S, McNamee R, Vail A. Bias in frequently reported analyses of subfertility trials. *Stat Med* 2008;27: 5605-5619.

Diggle P, Liang K-Y, Zeger SL. *Analysis of longitudinal data*, 1994. Clarendon Press ;

Oxford University Press, Oxford, New York.

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

Dunson DB. Bayesian latent variable models for clustered mixed outcomes. J R Stat Soc Series B Stat Methodol 2000;62: 355-366.

Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* 2003;59: 521-530.

Ernst E, Pittler MH. Efficacy or effectiveness? *J Intern Med* 2006;260: 488-490.

Ferraretti AP, Gianaroli L. The Bologna criteria for the definition of poor ovarian responders: is there a need for revision? *Hum Reprod* 2014;29: 1842-1845.

Fiedler K, Ezcurra D. Predicting and preventing ovarian hyperstimulation syndrome (OHSS): the need for individualized not standardized treatment. *Reprod Biol Endocrinol* 2012;10.

Fox J. Teacher's corner: structural equation modeling with the sem package in R. *Struct Equ Modeling* 2006;13: 465-486.

Gallot V, da Silva ALB, Genro V, Grynberg M, Frydman N, Fanchin R. Antral follicle responsiveness to follicle-stimulating hormone administration assessed by the Follicular Output RaTe (FORT) may predict in vitro fertilization-embryo transfer outcome. *Hum Reprod* 2012;27: 1066-1072.

Gameiro S, Boivin J, Peronace L, Verhaak CM. Why do patients discontinue fertility treatment? A systematic review of reasons and predictors of discontinuation in fertility treatment. *Hum Reprod Update* 2012;18: 652-669.

Garrido N, Bellver J, Remohi J, Simon C, Pellicer A. Cumulative live-birth rates per total number of embryos needed to reach newborn in consecutive in vitro fertilization (IVF) cycles: a new approach to measuring the likelihood of IVF success. *Fertil Steril* 2011;96: 40-46.

Gaskins AJ, Afeiche MC, Wright DL, Toth TL, Williams PL, Gillman MW, Hauser R, Chavarro JE. Dietary Folate and Reproductive Success Among Women Undergoing Assisted Reproduction. *Obstet Gynecol* 2014;124: 801-809.

Gelbaya TA, Tsoumpou I, Nardo LG. The likelihood of live birth and multiple birth after single versus double embryo transfer at the cleavage stage: a systematic review and meta-analysis. *Fertil Steril* 2010;94: 936-945.

Germond M, Urner F, Chanson A, Primi M-P, Wirthner D, Senn A. What is the most relevant standard of success in assisted reproduction? The cumulated singleton/twin delivery rates per oocyte pick-up: the CUSIDERA and CUTWIDERA. *Hum Reprod* 2004;19: 2442-2444.

Gnoth C, Maxrath B, Skonieczny T, Friol K, Godehardt E, Tigges J. Final ART success rates: a 10 years survey. *Hum Reprod* 2011;26: 2239-2246.

Goldstein H. Multivariate Multilevel Data *Multilevel Statistical Models*. 2003, pp. 139-146.

Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. *Stat Model* 2009;9: 173-197.

Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10: 37-48.

Griesinger G, Dafopoulos K, Schultze-Mosgau A, Felberbaum R, Diedrich K. What is the most relevant standard of success in assisted reproduction? Is BESST (birth emphasizing a successful singleton at term) truly the best? *Hum Reprod* 2004;19: 1239-1241.

Guerif F, Lemseffer M, Couet ML, Gervereau O, Ract V, Royere D. Serum antimullerian hormone is not predictive of oocyte quality in vitro fertilization. *Ann Endocrinol-Paris* 2009;70: 230-234.

Harrell F. Comment on 'Continuation ratio and cumulative proportional odds model'. https://stats.stackexchange.com/questions/92371/continuation-ratio-and-cumulative-proportional-odds-model. Accessed September 2017.

Heijnen E, Macklon NS, Fauser B. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod* 2004;19: 1936-1938.

Hirst WM, Vail A, Brison DR, Roberts SA. Prognostic factors influencing fresh and frozen IVF outcomes: an analysis of the UK national database. *Reprod Biomed Online* 2011;22: 437-448.

Hogan JW, Blazar AS. Hierarchical logistic regression models for clustered binary outcomes in studies of IVF-ET. *Fertil Steril* 2000;73: 575-581.

Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Stat Med* 1997;16: 239-257.

Hogan JW, Scharfstein DO. Estimating causal effects from multiple cycle data in studies of in vitro fertilization. *Stat Methods Med Res* 2006;15: 195-209.

Hojgaard A, Ottosen LDM, Kesmodel U, Ingerslev HJ. Patient attitudes towards twin pregnancies and single embryo transfer - a questionnaire study. *Hum Reprod* 2007;22: 2673-2678.

Human Fertilisation and Embryology Authority. Code of Practice. https://www.hfea.gov.uk/code-of-practice/32. Accessed September 2017.

Human Fertilisation and Embryology Authority. Fertility treatment 2014: Trends and figures. 2016.

Jackson D, White IR, Seaman S, Evans H, Baisley K, Carpenter J. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Stat Med* 2014;33: 4681-4694.

Jayaprakasan K, Hopkisson J, Campbell B, Johnson I, Thornton J, Raine-Fenning N. A randomised controlled trial of 300 versus 225 IU recombinant FSH for ovarian stimulation in predicted normal responders by antral follicle count. *Bjog-Int J Obstet Gy* 2010;117: 853-862.

Jonsdottir I, Lundin K, Bergh C. Double embryo transfer gives good pregnancy and live birth rates in poor responders with a modest increase in multiple birth rates: results from an observational study. *Acta Obstet Gynecol Scand* 2011;90: 761-766.

Kailasam C, Keay SD, Wilson P, Ford WCL, Jenkins JM. Defining poor ovarian response during IVF cycles, in women aged < 40 years, and its relationship with treatment outcome. *Hum Reprod* 2004;19: 1544-1547.

La Marca A, Sunkara SK. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Hum Reprod Update* 2014;20: 124-140.

Legro RS, Wu XK, Barnhart KT, Farquhar C, Fauser BCJM, Mol B, Conference HC, Comm S. Improving the Reporting of Clinical Trials of Infertility Treatments (IMPRINT): modifying the CONSORT statement. *Hum Reprod* 2014;29: 2075-2082.

Lintsen AME, Eijkemans MJC, Hunault CC, Bouwmans CAM, Hakkaart L, Habbema JDF, Braat DDM. Predicting ongoing pregnancy chances after IVF and ICSI: a national prospective study. *Hum Reprod* 2007;22: 2455-2462.

Luk J, Arici A. Does the ovarian reserve decrease from repeated ovulation stimulations? *Curr Opin Obstet Gynecol* 2010;22: 177-182.

Luke B, Brown MB, Wantman E, Baker VL, Grow DR, Stern JE. Second try: who returns for additional assisted reproductive technology treatment and the effect of a prior assisted reproductive technology birth. *Fertil Steril* 2013;100: 1580-1584.

Luke B, Brown MB, Wantman E, Lederman A, Gibbons W, Schattman GL, Lobo RA, Leach RE, Stern JE. Cumulative birth rates with linked assisted reproductive technology cycles. *N Engl J Med* 2012;366: 2483-2491.

Maheshwari A, McLernon D, Bhattacharya S. Cumulative live birth rate: time for a consensus? *Hum Reprod* 2015;30: 2703-2707.

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.

Marcus D, Marcus A, Johnson A, Marcus S. Infertility treatment: When is it time to give up? An internet-based survey. *Hum Fertil* 2011;14: 29-34.

McLernon DJ, Harrild K, Bergh C, Davies MJ, de Neubourg D, Dumoulin JC, Gerris J, Kremer JA, Martikainen H, Mol BW *et al.* Clinical effectiveness of elective single versus double embryo transfer: meta-analysis of individual patient data from randomised trials. *Bmj* 2010;341: c6945.

McLernon DJ, Maheshwari A, Lee AJ, Bhattacharya S. Cumulative Live Birth Rates After 1 or More Complete Cycles of IVF: A Population-Based Study of Linked Cycle Data from 178,898 Women. *Obstet Gynecol Surv* 2016a;71: 290-291.

McLernon DJ, Steyerberg EW, te Velde ER, Lee AJ, Bhattacharya S. Predicting the chances of a live birth after one or more complete cycles of in vitro fertilisation: population based study of linked cycle data from 113 873 women. *Bmj* 2016b;355.

Min JK, Breheny SA, MacLachlan V, Healy DL. What is the most relevant standard of success in assisted reproduction? The singleton, term gestation, live birth rate per cycle initiated: the BESST endpoint for assisted reproduction. *Hum Reprod* 2004;19: 3-7.

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.

Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *Bmj* 2010;340.

Mohiyiddeen L, Newman WG, Cerra C, Horne G, Mulugeta B, Byers H, Roberts SA, Nardo LG. FSH receptor genotype does not predict metaphase-II oocyte output or fertilization rates in ICSI patients. *Reprod Biomed Online* 2013a;27: 305-309.

Mohiyiddeen L, Newman WG, Cerra C, McBurney H, Mulugeta B, Roberts SA, Nardo LG. A common Asn680Ser polymorphism in the follicle-stimulating hormone receptor gene is not associated with ovarian response to gonadotropin stimulation in patients undergoing in vitro fertilization. *Fertil Steril* 2013b;99: 149-155.

Moragianni VA, Penzias AS. Cumulative live-birth rates after assisted reproductive technology. *Curr Opin Obstet Gynecol* 2010;22: 189-192.

Nelson SM, Lawlor DA. Predicting Live Birth, Preterm Delivery, and Low Birth Weight in Infants Born from In Vitro Fertilisation: A Prospective Study of 144,018 Treatment Cycles. *PLoS Med* 2011;8.

NHS Choices. Infertility.http://www.nhs.uk/conditions/Infertility/Pages/Introduction.aspx Last accessed September 2017.

Nyboe Andersen A, Nelson SM, Fauser B, Garcia-Velasco JA, Klein BM, Arce JC. Individualized versus conventional ovarian stimulation for in vitro fertilization: A multicenter, randomized, controlled, assessor-blinded, phase 3 noninferiority trial. *Fertil Steril* 2017;107: 387-396.

Olivius C, Friden B, Borg G, Bergh C. Why do couples discontinue in vitro fertilization treatment? A cohort study. *Fertil Steril* 2004;81: 258-261.

Olivius K, Friden B, Lundin K, Bergh C. Cumulative probability of live birth after three in vitro fertilization/intracytoplasmic sperm injection cycles. *Fertil Steril* 2002;77: 505-510.

Pearl J. Causal diagrams for empirical research. *Biometrika* 1995;82: 669-688.

Pelinck M, Vogel N, Arts E, Simons A, Heineman M, Hoek A. Cumulative pregnancy rates after a maximum of nine cycles of modified natural cycle IVF and analysis of patient drop-out: a cohort study. *Hum Reprod* 2007;22: 2463-2470.

Penman R, Heller G, Tyler J. Modelling IVF Data using an Extended Continuation Ratio Random Effects Model *Proceedings of the 22nd International Workshop on Statistical Modelling*. 2007, Barcelona.

Pinborg A, Loft A, Ziebe S, Andersen AN. What is the most relevant standard of success in assisted reproduction? Is there a single 'parameter of excellence'? *Hum Reprod* 2004;19: 1052-1054.

Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modeling. *Psychometrika* 2004;69: 167-190.

Rizopoulos D. *Joint models for longitudinal and time-to-event data: With applications in R*, 2012. CRC Press.

Roberts S, Hirst W, Brison D, Vail A, Collaboration t. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. *Hum Reprod* 2010a;25: 2792-2802.

Roberts SA. Models for assisted conception data with embryo-specific covariates. *Stat Med* 2007;26: 156-170.

Roberts SA, McGowan L, Hirst WM, Brison DR, Vail A, Lieberman BA. Towards single embryo transfer? Modelling clinical outcomes of potential treatment choices using multiple data sources: predictive models and patient perspectives. *Health Technol Assess* 2010b;14: 1-+.

Roberts SA, Stylianou C. The non-independence of treatment outcomes from repeat IVF cycles: estimates and consequences. *Hum Reprod* 2012;27: 436-443.

Robins JM. Marginal structural models versus structural nested models as tools for causal inference *Statistical models in epidemiology, the environment, and clinical trials*. 2000. Springer, pp. 95-133.

Rubin DB. Inference and Missing Data. *Biometrika* 1976;63: 581-590.

Schieve LA, Reynolds MA. What is the most relevant standard of success in assisted reproduction? Challenges in measuring and reporting success rates for assisted reproductive technology treatments: What is optimal? *Hum Reprod* 2004;19: 778-782.

Sharif K, Afnan M. The IVF league tables: time for a reality check. *Hum Reprod* 2003;18: 483-485.

Society for Assisted Reproductive Technologies. National Summary Report 2015. 2016.

Soullier N, Bouyer J, Pouly JL, Guibert J, de La Rochebrochard E. Estimating the success of an in vitro fertilization programme using multiple imputation. *Hum Reprod* 2008;23: 187-192.

Stern JE, Brown MB, Luke B, Wantman E, Lederman A, Missmer SA, Hornstein MD. Calculating cumulative live-birth rates from linked cycles of assisted reproductive technology (ART): data from the Massachusetts SART CORS. *Fertil Steril* 2010;94: 1334-1340.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* 2009;339.

Stewart LM, Holman CAJ, Hart R, Finn J, Mai Q, Preen DB. How effective is in vitro fertilization, and how can it be improved? *Fertil Steril* 2011a;95: 1677-1683.

Stewart LM, Holman CD, Hart R, Finn J, Mai Q, Preen DB. How effective is in vitro fertilization, and how can it be improved? *Fertil Steril* 2011b;95: 1677-1683.

Stillman RJ, Richter KS, Jones HW, Jr. Refuting a misguided campaign against the goal of single-embryo transfer and singleton birth in assisted reproduction. *Hum Reprod* 2013;28: 2599-2607.

Sundstrom P, Saldeen P. Cumulative delivery rate in an in vitro fertilization program with a single embryo transfer policy. *Acta Obstet Gynecol Scand* 2009;88: 700-706.

Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. *Hum Reprod* 2011;26: 1768-1774.

Templeton A, Morris JK, Parslow W. Factors that affect outcome of in-vitro fertilisation treatment. *Lancet* 1996;348: 1402-1406.

Ten Have TR, Kunselman AR, Pulkstenis EP, Landis JR. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics* 1998;54: 367-383.

Troude P, Guibert J, Bouyer J, de La Rochebrochard E, Grp D. Medical factors associated with early IVF discontinuation. *Reprod Biomed Online* 2014;28: 321-329.

Vail A, Gardener E. Common statistical errors in the design and analysis of subfertility trials. *Hum Reprod* 2003;18: 1000-1004.

Van Voorhis BJ. Clinical practice. In vitro fertilization. *The N Engl J Med* 2007;356: 379-386.

Veleva Z, Karinen P, Tomas C, Tapanainen JS, Martikainen H. Elective single embryo transfer with cryopreservation improves the outcome and diminishes the costs of IVF/ICSI. *Hum Reprod* 2009;24: 1632-1639.

Verberg MF, Eijkemans MJ, Heijnen EM, Broekmans FJ, de Klerk C, Fauser BC, Macklon NS. Why do couples drop-out from IVF treatment? A prospective cohort study. *Hum Reprod* 2008;23: 2050-2055.

Verhagen TE, Dumoulin JC, Evers JL, Land JA. What is the most accurate estimate of pregnancy rates in IVF dropouts? *Hum Reprod* 2008;23: 1793-1799.

Witsenburg C, Dieben S, Van der Westerlaken L, Verburg H, Naaktgeboren N. Cumulative live birth rates in cohorts of patients treated with in vitro fertilization or intracytoplasmic sperm injection. *Fertil Steril* 2005;84: 99-107.

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

II. Measuring multistage IVF outcomes

Chapter 2. Methods for reviews of outcome measurements in use in IVF

Our reviews of outcome measures used in IVF are presented as journal articles 1,2, and 3 (Chapters 3, 4 and 5). Each of these journal articles contains methods sections, articulating the factual details of the searches, data extraction and analysis. To avoid repetition and tedium, the following methods chapter will focus on the thinking behind the approaches employed in those articles.

2.1 Motivation

What outcomes are actually used for evaluation of IVF treatments? A review of the outcome measures reported for IVF would be informative. Specifically, this would indicate the reporting standards in the field and would identify the principle treatment stages to include in our models. We have already seen that different outcome measures will be needed in different contexts. Although our intention is to focus on models for mechanism and efficacy, clinical endpoints remain within our scope; we would be interested in investigating how procedural outcomes of treatment relate to birth outcomes, for example. Moreover, even in settings where the primary audience is prospective or current patients, we can get a sense of how clinicians partition the cycle into chunks by looking at the variety of denominators they use (an outcome reported 'per transfer' represents a severing of the cycle at the transfer procedure, for example). On the basis of these considerations, we decided that it would be appropriate to review outcome reporting in two rather different settings. First, we conducted a review of outcome reporting in RCTs of IVF. Second, we looked at reporting of success rates on IVF clinic websites, which are targeted at prospective patients. In the case of RCTs, we included both efficacy and effectiveness trials, so as to capture as much of the variation in outcome reporting as possible.

2.1.1. The relevance of core outcome sets

Concern over the lack of consistency in reported outcome measures in women's health research has led to the formation of the CROWN (CoRe Outcomes in WomeN's health) Initative, comprising the editors of over 50 women's health journals (Khan, 2014). CROWN has noted that the lack of core outcome sets in women's health research hinders comparisons between studies and the synthesis of evidence, and have called for efforts to plug the gap. A systematic review of primary outcomes reported in trials and systematic reviews of interventions for preventing preterm birth has been given as an example; the review authors found seventy-two different primary outcomes reported in 103 RCTs and 29 different primary outcomes in 33 Cochrane reviews (Meher and Alfirevic, 2014). In that review (of trials and Cochrane reviews), the authors noted the need for reviews of outcomes within each specialty as a precursor to the development of a core outcome set. Although it is not our concern to establish a core outcome set for infertility research, a review of outcomes used in IVF research would constitute a useful contribution to this ongoing endeavour.

2.2 Redundancy check in literature databases

As has been noted above, the appropriate choice of outcome is likely to be context dependent, so that different endpoints might be suitable for summarising clinic performance and for research. Regardless of suitability, we might expect that different outcomes might be used in each of these domains. A review of IVF outcomes should distinguish between these settings by reporting on both separately. In the remainder of this section, I will outline the considerations for conducting reviews within both of these domains. In order to preclude duplication of work that may have been completed or ongoing elsewhere, a search was conducted in the MEDLINE and COMET (Core Outcome Measures in Effectiveness Trials) databases. The search string used in a MEDLINE search is shown in 2.7. In COMET, searches can be made by condition. Searches were made for each of ‘infertility’ and ‘subfertility’. The COMET search returned one article (Dapuzzo, et al., 2011). The MEDLINE search returned 32 articles; abstracts and titles were screened and one was considered to be directly relevant (Kushnir, et al., 2013). A search for articles citing either of these was made in the Web of Science database. None of the citing articles constituted reviews of outcomes in use.

The first of the two identified articles was a review of outcomes reported in 294 RCTs published between 2004 and 2010 in the top 10 Obstetrics and Gynecology journals according to 2008 ISI Rankings (Dapuzzo, et al., 2011). The authors only included trials with pregnancy as one of the outcomes. Furthermore, they restricted their focus to outcomes relating to pregnancy (including live birth), pregnancy loss and adverse events. Around 40% of the included trials reported on live birth. The second article attempted to assess the transparency of reporting by clinics using the surveillance reports of the Society for Assisted Reproductive Technologies and Centres for Disease Control and Prevention. The authors compared the numbers of started cycles with the numbers of cycles for which outcomes were reported, and noted an increasing proportion of excluded cycles between 2005 and 2010 (Kushnir, et al., 2013). The analysis may be critically flawed however, as the authors appear to have included embryo-banking cycles, where stimulation is performed with the intention of freezing embryos for future use (Kissin, et al., 2013). Increasing use of embryo-banking cycles enabled by improvements in cryopreservation technology may therefore account for much of this trend.

Clearly, neither of these articles would render our review (really, reviews) redundant. In the remainder of this section, I will outline the relevant considerations for reviews of reported outcomes in IVF research and on IVF clinic websites.

2.3 Methodological considerations for a review of outcome measures in IVF research

2.3.1. Which studies should we include?

While we noted above that we believe our scope should be fairly broad, so as to capture response measures employed at all different stages and to different ends, some narrowing down of the eligibility criteria from ‘any IVF research’ would clearly be appropriate. One way to reduce the number of eligible studies would be to focus solely on randomised trials. This might be justified by noting that ‘observational study’ is a vague designation, and would potentially open the door to pieces of work that are not obviously research. The line between a routine audit and a retrospective cohort study might turn out to be thin. Moreover, whereas for trials, the population of studies is delineated by a well-understood and quite clearly defined class of study designs, it is not clear which research designs should be eligible in the observational scenario. This is largely because this group of studies is defined by the lack of allocation of groups to treatment by researchers and therefore incorporates a heterogeneous collection of designs.

The decision to review outcome measures in IVF RCTs prompts the question of which RCTs to include. For systematic reviews of interventions, this entails deciding upon a definition for the intervention, or interventions, and a patient population. Randomised controlled trials testing this intervention in the chosen population are then eligible for inclusion in the review. In the present case, interest lies in the outcomes that are in use in trials in IVF, and the relevant considerations may differ slightly compared to a review of a particular intervention. If all trials of interventions for subfertile patients were considered eligible, then this would include alternative but sometimes similar treatments, such as intrauterine insemination (IUI), where sperm is selected and placed into the uterus during the ovulation period, and intra-cytoplasmic sperm injection (ICSI), which differs from usual IVF in that the sperm is injected directly into the egg to create the embryo. Arguments for including these treatments include general coincidence of purpose and patient populations, with many clinics offering these as alternatives to IVF. Indeed, ICSI is often considered as a variant of IVF and ICSI cycles are often included in analyses of IVF programmes. Arguments against include the fact that some of these treatments will be too dissimilar to be directly relevant to the present project and the fact that increasing the scope of the review may render the exercise infeasible. A solution may lie in prespecifying those treatments that are sufficiently similar to be considered eligible (so that, for example, trials in patients undergoing ICSI may be included, but trials in patients undergoing IUI may not). Such decisions must be to some extent arbitrary. Similar decisions must be made in relation to trials of interventions delivered alongside IVF as opposed to interventions consisting of modifications to the process itself. For example, a recent systematic review of psychological and educational interventions for subfertile men and women included trials in which patients were given some psychological intervention before, during or after IVF (Verkuijlen, et al., 2016). Although this does not constitute direct modification of the IVF process, it can be argued that the outcomes reported in such trials are relevant both to patients undergoing IVF and to the present review. Moreover, in some cases, the matter of whether an intervention constitutes a supplement or a modification to the IVF treatment may be unclear.

It would be necessary to define a time period during which a study must have been published in order to be eligible. A key reason for this is again feasibility of conducting the review. However, it can be argued that a review of more recent studies may be more informative and relevant on two grounds. Firstly, we anticipate that recent trials will be of a higher standard in comparison to earlier work. A review of subfertility RCTs published in 2001 highlighted poor standards of study design and reporting (Vail and Gardener, 2003). A review of subfertility RCTs published in 1990, 1996 and 2002 also highlighted the historically poor methodological quality and standard of reporting in this field, although the authors noted that there was evidence of improvement over the period (Dias, et al., 2006). However, methodological quality might not matter in the present review (see 2.3.2, below). Furthermore, poor standards of outcome reporting would constitute a valuable finding in this context. A second argument for focussing on recent studies is that much of the debate around the appropriate outcome for IVF research was triggered by the proposal of BESST in 2004 (Min, et al., 2004). Research published following this discussion is more likely to reflect the current trends relating to outcome reporting, which are the focus of the review.

2.3.2. What should we extract?

A review of outcomes would differ from a review of trials of an intervention in that, in the case of the former, there is no interest in estimating the effectiveness or safety of a treatment through data synthesis. The methodological quality of the studies is not directly relevant to the question of what outcomes are reported, so no assessment of risk of bias in the included studies would need to be made. The focus should be on extracting the outcomes that are actually reported in the trials. Outcome definitions should include both the numerators and denominators used in the study. Some additional characteristics of the trials should be extracted so that results can be presented separately for studies with different features (which is not to say that any inferential comparisons are intended). Characteristics that could be extracted include information about the patients, intervention, setting, objectives, study design, duration, sample size, rates of attrition, analysis, funding source and the actual results of the trial. It might be necessary to prune this list to a relevant subset for the purposes of practicability.

2.3.3. **Search strategy**

Although the interest in Cochrane Reviews is usually in evaluation of an intervention, the general search strategy recommended by Cochrane might also be suitable for a review of outcomes. In a Cochrane review, the central concern of the search strategy is to minimise publication bias arising from the fact that ease of access to trial reports is related to their results (Higgins and Green, 2011). For a review of reported outcomes, the concern is not related to bias but rather to ensuring that the findings are representative of and generalisable to the population of trials defined by the review inclusion criteria. It is not clear whether or not ease of access to trial reports is related to the outcomes reported within, although such a relationship would be plausible due to within-study selective reporting of significant results. However, an exhaustive search would protect against both sources of concern by identifying all trials in the population defined by the eligibility criteria of the review. Given that this is precisely what the Cochrane search strategy is intended to do, it would appear to be reasonable to employ a similar search methodology.

Briefly, the search strategy recommended by Cochrane consists of a search of large databases of trials (MEDLINE, EMBASE, CENTRAL) supplemented by searches of specialised registers of trials and of the grey literature. In MEDLINE, searches should be highly sensitive (in the sense of returning large numbers of potentially relevant studies) even at the expense of precision (in the sense that very few of the studies returned by the search will actually be eligible for inclusion). The reasoning is that it is important to capture all relevant studies and that the time taken to screen a study for eligibility is short (possibly as low as 30 seconds) (Lefebvre, et al., 2008).

Using the published search strategies in systematic reviews of the Cochrane Gynaecology and Fertility Editorial Group as a guide, MEDLINE search strings were constructed and piloted in order to explore the number of trials that would have to be screened initially, and therefore the feasibility of the review. The strings were designed to be broad enough to capture assisted reproductive technologies other than IVF, and explicitly included terms relating to IUI and ICSI. The searches are shown in 2.7. The MEDLINE search returned 2621 results, of which 1422 were published in the last 10 years. Assuming a rate of screening of 30 seconds per returned study, this translates to around 12 hours of screening. It is unclear exactly how many of the studies would be eligible for subsequent data extraction; this would depend on the exact inclusion criteria used. Searches of other databases were not conducted as part of this initial scoping exercise.

2.4 Methodological considerations for a review of outcomes reported by IVF clinics

2.4.1. Inclusion criteria

We would like to know the outcomes reported by IVF clinics in the UK. As such, inclusion criteria in this case could be defined as UK clinics performing IVF (or possibly related treatments such as ICSI). In relation to this and other aspects, a review of outcomes reported by clinics should be more straightforward than the corresponding review of trials.

2.4.2. What should we extract?

In relation to clinics, possible information to extract includes the outcomes reported (numerator and denominator), the results themselves, the sample sizes and stratification variables (potentially including time periods and results for different treatments).

2.4.3. Search strategy

Particular interest lies in the outcomes that clinics use in direct communication with prospective and existing patients, rather than in the outcomes that are presented collectively by HFEA. This information is held on the websites of clinics. These websites were listed on the HFEA's ‘Choose a Fertility Clinic’ database[2]. A search of this database revealed 72 clinics in the UK offering IVF or ICSI (one clinic offers IVF but not ICSI, no clinics offer ICSI but not IVF). A review of outcomes reported by clinic websites would not therefore be prohibitively difficult.

[2] This website has now been superseded by the new *Choose a Clinic* website https://www.hfea.gov.uk/choose-a-clinic/

2.5 Summary of Chapter 2.

Although a review of outcomes used in IVF would seem to be warranted, considerable thought needs to be given to the matter of what kinds of studies to incorporate. A key judgement relates to whether or not it is essential or even desirable to include hard to find studies in order to answer the review question. Aside from this, it may not be feasible to conduct the review of research sketched above due to the large numbers of studies that would need to be screened and subsequently interrogated for relevant data. Feasibility of the exercise could be improved by reducing the scope of the review. For example, the review of trials could be dropped, on the grounds that a similar review was conducted in 2011 (Dapuzzo, et al., 2011). The focus on pregnancy and live birth outcomes, the failure to report on denominators and the restriction to the top 10 journals may be seen as notable limitations of that study however. Another key decision is whether to include studies of treatments related or similar to IVF (such as IUI, ICSI) and of treatments delivered as supplementary to IVF. A review of outcomes reported on the websites of UK clinics promises to be more straightforward, with a clearly defined population and free access to the required information.

In the remainder of Part II, we present three journal articles on the topic of outcome measures in IVF. The first two of these comprise the reviews of trials and clinic websites described above. In the third, we challenge the decision of HFEA to switch the performance indicator they present to patients on their website to the measure 'live birth per embryo transferred'. As a debate article, there are no methods to report as such. The objections we present are based on arguments formulated during our review studies. We suggest that no single measure is a sufficient performance indicator, and that a suite of measures should be reported instead.

2.6 References for Chapter 2.

Dapuzzo L, Seitz FE, Dodson WC, Stetter C, Kunselman AR, Legro RS. Incomplete and inconsistent reporting of maternal and fetal outcomes in infertility treatment trials. *Fertil Steril* 2011;95: 2527-2530.

Dias S, McNamee R, Vail A. Evidence of improving quality of reporting of randomized controlled trials in subfertility. *Hum Reprod* 2006;21: 2617-2627.

Higgins JP, Green S. *Cochrane handbook for systematic reviews of interventions*, 2011. John Wiley & Sons.

Khan K. The CROWN Initiative: Journal editors invite researchers to develop core outcomes in women's health. *Midwifery* 2014;30: 1147-1148.

Kissin DM, Crawford S, Boulet SL. The status of public reporting of clinical outcomes in assisted reproductive technology. *Fertil Steril* 2013;100: E16-E17.

Kushnir VA, Vidali A, Barad DH, Gleicher N. The status of public reporting of clinical outcomes in assisted reproductive technology. *Fertil Steril* 2013;100: 736-+.

Lefebvre C, Manheimer E, Glanville J. Searching for studies. *Cochrane handbook for systematic reviews of interventions: Cochrane book series* 2008: 95-150.

Meher S, Alfirevic Z. Choice of primary outcomes in randomised trials and systematic reviews evaluating interventions for preterm birth prevention: a systematic review. *Bjog-Int J Obstet Gy* 2014;121: 1188-1194.

Min JK, Breheny SA, MacLachlan V, Healy DL. What is the most relevant standard of success in assisted reproduction? The singleton, term gestation, live birth rate per cycle initiated: the BESST endpoint for assisted reproduction. *Hum Reprod* 2004;19: 3-7.

Vail A, Gardener E. Common statistical errors in the design and analysis of subfertility trials. *Hum Reprod* 2003;18: 1000-1004.

Verkuijlen J, Verhaak C, Nelen WLDM, Wilkinson J, Farquhar C. Psychological and educational interventions for subfertile men and women. *Cochrane Db Syst Rev* 2016.

2.7 Supplementary Material for Chapter 2.

Search string used for redundancy search in MEDLINE

Keywords CONTAINS “IVF” or “in vitro fertilization” or “in-vitro fertilisation” or “ICSI” or “intracytoplasmic sperm injection” or “Embryo” or “in-vitro fertilization” or “ART” or “assisted conception” or “assisted reproduction” or “artificial insemination” or “IUI” or “IVF-ET” or “subfertility” or “Infertility” or Title CONTAINS “IVF” or “in vitro fertilization” or “in-vitro fertilisation” or “ICSI” or “intracytoplasmic sperm injection” or “Embryo” or “in-vitro fertilization” or “ART” or “assisted conception” or “assisted reproduction” or “artificial insemination” or “IUI” or “IVF-ET” or “subfertility” or “Infertility” AND

Keywords CONTAINS core outcome or core outcomes or outcome reporting or Title CONTAINS core outcome or core outcomes or outcome reporting

Search string used to investigate approximate number of eligible trials in MEDLINE

Search (((((“ovarian stimulation”[MeSH Terms] OR “ovarian hyperstimulation”[MeSH Terms])) OR (“ovarian stimulation”[Title] OR “ovarian hyperstimulation”[Title]))) OR (((“IVF”[MeSH Terms] OR “in vitro fertilization”[MeSH Terms] OR “in-vitro fertilisation”[MeSH Terms] OR “ICSI” or“intracytoplasmic sperm injection”[MeSH Terms] OR “Embryo”[MeSH Terms] OR “in-vitro fertilization”[MeSH Terms] OR “ART”[MeSH Terms] OR “assisted conception”[MeSH Terms] OR “assisted reproduction”[MeSH Terms] OR “artificial insemination”[MeSH Terms] OR “IUI”[MeSH Terms] OR “IVF-ET”[MeSH Terms] OR “subfertility”[MeSH Terms] OR “Infertility”[MeSH Terms])) OR (“IVF”[Title] OR “in vitro fertilization”[Title] OR “in-vitro fertilisation”[Title] OR “ICSI” or“intracytoplasmic sperm injection”[Title] OR “Embryo”[Title] OR “in-vitro fertilization”[Title] OR “ART”[Title] OR “assisted conception”[Title] OR “assisted reproduction”[Title] OR “artificial insemination”[Title] OR “IUI”[Title] OR “IVF-ET”[Title] OR “subfertility”[Title] OR “Infertility”[Title])))) AND (((randomised controlled trial[MeSH Terms] OR randomized controlled trial[MeSH Terms] OR randomized[MeSH Terms] OR randomised[MeSH Terms]

OR placebo[MeSH Terms] OR clinical trials as topic[MeSH Terms] OR randomly[MeSH Terms] OR trial[MeSH Terms] OR crossover[MeSH Terms] OR cross-over[MeSH Terms] OR cross over[MeSH Terms])) OR (randomised controlled trial[Title] OR randomized controlled trial[Title] OR randomized[Title] OR randomised[Title] OR placebo[Title] OR clinical trials as topic[Title] OR randomly[Title] OR trial[Title] OR crossover[Title] OR cross-over[Title] OR cross over[Title])).

Chapter 3. Direct to consumer advertising of success rates for medically assisted reproduction: a review of national clinic websites.

Journal article 1

Authors Jack Wilkinson, Andy Vail, Stephen A Roberts

Status Published in BMJ Open

Reference Wilkinson, J., Vail, A., Roberts, S.A. (2017). "Direct-to-consumer advertising of success rates for medically assisted reproduction: a review of national clinic websites." BMJ Open **7**(1): e012218.

Contribution statement JW designed the study, and undertook the acquisition, analysis and interpretation of data, drafted the manuscript and gave final approval of the version to be published. All other authors contributed to the design of the study, the interpretation of data, drafting and revision of the manuscript and gave final approval of the version to be published. JW is acting as guarantor for the study.

Preamble Our primary motivation for reviewing the IVF outcomes in use was to inform subsequent method development (specifically, to identify the stages and interventions to include in our models). However, it became apparent during this review of clinic websites that the multiplicity of potential outcome measures in IVF could be abused when employed for the purposes of marketing a clinic's services. Accordingly, we felt that it was important to report the outcome heterogeneity we found and to describe the implications for consumers.

Outputs and Impact of the research Oral presentations relating to this work were delivered at the University of Manchester Institute for Population Health Student Showcase, the Cochrane Gynaecology and Fertility 20 year anniversary conference, and by invitation to a meeting of the North of England Reproductive Medicine Group, and at an internal seminar of Central Manchester NHS Foundation Trust Department of Reproductive Medicine. In addition, the study received some media coverage, appearing in national newspapers (eg: The Daily Mail, The Telegraph, The Sun, The Independent). JW was interviewed for Irish national radio (Newstalk) and local television (That's Manchester) about the study. On the basis of this work, JW was invited by the Human

Fertilization and Embryology Authority (HFEA) to discuss their plans for reporting of IVF success rates on their website. At the time of writing, the authors have received assurances from HFEA that they will address the issues raised in the study.

Finally, the publication of the study has garnered interest from patient representatives, particularly on social media. This has led to an invitation to JW to act as an advisor on a new patient information initiative (www.ReproTechTruths.org).

3.1 Abstract

Objectives

To establish how medically assisted reproduction (MAR) clinics report success rates on their websites.

Setting

Websites of private and NHS clinics offering in vitro fertilisation in the United Kingdom.

Participants

We identified clinics offering in vitro fertilisation (IVF) using the Choose a Fertility Clinic facility on the website of the Human Fertilisation and Embryology Authority (HFEA). Of 81 clinics identified, a website could not be found for two, leaving 79 for inclusion in the analysis.

Primary and secondary outcome measures

Outcome measures reported by clinic websites. Both the numerator and denominator included in the outcome measure were of interest.

Results

53 (67%) websites reported their performance using 51 different outcome measures. It was most common to report pregnancy (83% of these clinics) or live birth rates (51%). Thirty-one different ways of reporting pregnancy and nine different ways of reporting live birth were identified. Eleven (21%) reported multiple birth or pregnancy rates. One clinic provided information on adverse events. It was usual for clinics to present results without relevant contextual information such as sample size, reporting period, the characteristics of patients, and particular details of treatments.

Conclusions

Many combinations of numerator and denominator are available for the purpose of reporting success rates for MAR. The range of reporting options available to clinics is further increased by the possibility of presenting results for subgroups of patients and for different time periods. Given the status of these websites as advertisements to patients the risk of selective reporting is considerable. Binding guidance is required to ensure consistent, informative reporting.

KEYWORDS

In vitro fertilisation. Advertising. Success rates. Assisted reproduction. Live birth rate.

3.2 Introduction

Direct to consumer advertising of prescription drugs is permitted only in the United States and New Zealand. However, concerns that direct advertising drives demand for more expensive, rather than more effective, treatments do not extend to bans on direct advertising of other medical practices.

Questionnaires of subfertile patients have indicated that a majority make use of the internet to find information relating to their condition (Haagen, et al., 2003, Rawal and Haddad, 2006) with a recent survey in Poland suggesting that 93% of respondents used online resources for this purpose (Talarczyk, et al., 2012). A key decision for any patient seeking treatment for subfertility is where to be treated, and it is expected that patients will take performance into account when choosing a fertility clinic. In practice, the reporting of success rates for medically assisted reproduction (MAR) is complicated by the complex, multi-stage nature of the treatments involved. Taking an in vitro fertilization (IVF) cycle as an example, patients will typically undergo a period of ovarian stimulation before eggs are recovered and then fertilized. Some of the resulting embryos are then transferred to the uterus with the objectives of pregnancy and the subsequent birth of a healthy child. Failure may occur at each step in this sequence, so that a considerable variety of numerators (such as pregnancy or live birth) and denominators (such as started cycles, transfer procedures, or egg collections) may be used (Heijnen, et al., 2004). Furthermore, since patients typically undertake multiple attempts at treatment, there is the option to report outcomes in a cumulative fashion. For example, live birth rates could be reported following several stimulation or transfer procedures. Consequently, the matter of how MAR success rates should be reported has been extensively discussed in the literature (Abdalla, et al., 2010, Garrido, et al., 2011, Germond, et al., 2004, Griesinger, et al., 2004, Meldrum, 2013, Min, et al., 2004, Pinborg, et al., 2004, Schieve and Reynolds, 2004) and has featured in a recent consultation process ('Information for Quality') by the Human Fertilisation and Embryology Authority ((HFEA) 2014). There is also the question of how to report adverse consequences of treatment. In particular, given the HFEA policy of reducing the number of twin births arising from MAR, the reporting of multiple pregnancy rates requires attention.

In addition to informing clinic selection, reported outcomes may also be used by patients trying to understand their own chances of success. At present, HFEA present success rates in the form of live birth per cycle started and live birth per embryo transferred on its online Choose a Fertility Clinic facility, a new version of which is currently being tested (Human Fertilisation and Embryology Authority, 2009). This information is presented separately for treatments involving fresh and frozen embryos, for patients using their own or using donated gametes, and for different age groups. Furthermore, the particular treatment variants included in the results, the sample sizes, and the reporting period are all presented. In principle, the provision of this contextual information makes it possible for patients to identify relevant results and to consider these when making decisions about whether and where to commence treatment. Although HFEA provide standardised reporting of success rates, no such standardisation is imposed on clinics' own websites. In light of this, the consistency and clarity of online reporting is of material interest.

In order to investigate the standards of reporting of MAR success rates, we conducted a national review of MAR clinic websites. Our aim was to identify the outcomes in use by clinics and to examine whether results were presented in a consumer-friendly manner.

3.3 Methods

3.3.1. Identification of websites

We restricted our focus specifically to clinics offering assisted reproductive technology (ART), although we extracted information about other MAR treatments, such as intra-uterine insemination (IUI), which would not be considered ART (Zegers-Hochschild, et al., 2009). An initial search was made between 26/01/2015 and 29/01/15 on the HFEA Choose a Fertility Clinic facility (Human Fertilisation and Embryology Authority, 2009) using the search options ‘both’ for the field ‘funding for patients’ and ‘IVF’ for ‘treatments offered’. An earlier scoping exercise had suggested that no clinic offered intra-cytoplasmic sperm injection (ICSI) but not in-vitro fertilisation (IVF). This search was performed for each of the 12 ‘regions’ listed by HFEA. The website addresses of each clinic were recorded. Where the website listed by HFEA was inactive, or where no website was listed, the correct address was obtained via Google searching. It became apparent that this method had not produced a complete list of clinics. Accordingly, a further search was made using the A to Z listings on the HFEA website on 04/02/15 and 05/02/15. Any clinics offering IVF that were not identified during the initial search were added to the dataset. Again, missing or defunct website addresses were updated by searching on Google. As a final check, the initial search was repeated on 05/02/15 with the 'funding for patients' field replaced by each of 'private' and 'nhs'. Although this revealed clinics that had not been identified during the initial search, it did not reveal any clinics that had not been identified after the A to Z search. Where multiple clinics shared a website, we used the centre-specific results for analysis, so that the clinic was the unit of analysis.

3.3.2. Data extraction

Data were extracted at both the clinic-level and for each reported result on the clinic's website. At the clinic-level, we recorded the type of patients treated (NHS, private or both), whether or not an NHS logo was displayed on the front page, whether or not patient testimonials were used, and if so whether or not these were featured on the front page, whether selection policies relating to body mass index (BMI), age, number of previous attempts and smoking status were reported and whether the website reported success rates. At the result-level, we extracted the numerator and denominator used, together with the definition of the numerator if provided. We further extracted the corresponding patient and cycle characteristics for the reported item, including patient age range, treatments included, whether donor gametes were included, whether fresh or frozen cycles were included (for treatments other than intrauterine insemination (IUI)), the sample size, the reporting period as well as the number of cancellations and incomplete treatments. For each of these, we recorded instances where the required information could not be identified from the presented results.

3.3.3. Statistical analysis

We summarised the characteristics of the clinic websites, tabulating the numerators and denominators in use within five categories: pregnancy, live birth, multiple births, pre-clinical outcomes and adverse events. We calculated the proportion of clinics where results were presented in such a way so that each of age range, included treatments, inclusion of donor gametes, inclusion of fresh/ frozen cycles, sample size, number of abandoned treatments and reporting period could not be identified. We were particularly interested in whether or not clinics achieved the standard of reporting adopted by HFEA. To this end, we calculated the proportion of websites reporting the outcomes 'live birth per cycle started' and 'live birth per embryo transferred' together with all of the relevant contextual information (that is, all of the factors listed above with the exception of number of abandoned treatments, as these cycles are included as failures in rates reported per cycle started).

We calculated the proportion of websites for which patient selection policies were not stated. Finally, we made a tentative comparison between NHS and private clinics in relation to standards of reporting, although we did not consider statistical inference to be particularly meaningful in relation to this.

3.4 Results

3.4.1. Characteristics of clinics

The search identified 81 clinics in the UK. Of the 81 clinics identified, a website could not be found for two, leaving 79 for the present analysis. Fifty-three (67%) reported outcomes. Amongst those reporting outcomes, there was considerable variation in the number reported; the median (range) was 36 (1 to 127). Sixty-two (78%) stated that they treated both NHS and private patients, four (5%) described themselves as treating NHS patients only, and 13 (16%) stated that they exclusively treated private patients. Twenty-three (29%) displayed an NHS logo on the front page. Forty-nine (62%) of the websites featured patient testimonials, of which 23 (47%) featured these on the front page.

3.4.2. Reported outcomes

A total of 54 different outcome measures were identified during the review. The distribution of clinical outcome measures across the clinics is shown in Figure 4 and Figure 5.

Pregnancy outcomes

Thirty-three different ways of reporting pregnancy were identified (Table 1). The majority (81%) of clinics reported clinical pregnancy rates, with most (55%) websites reporting these per transfer procedure. A substantial proportion (36%), although fewer than half, reported clinical pregnancy per cycle started. Notably, around one in four websites reported clinical pregnancy rates without specifying the denominator. Just under a fifth (19%) of websites presented biochemical pregnancy rates, and these were most commonly reported per transfer (11%), per cycle started (8%), or without specifying the denominator (8%). Over a fifth (21%) of clinics presented pregnancy rates without explaining what was meant by ‘pregnancy’, with 15% also leaving the denominator unspecified. Reporting of cumulative outcomes across multiple transfers or inseminations was sparse, with no site reporting biochemical pregnancies and only a small number reporting clinical pregnancy rates cumulatively. One site reported continuing pregnancy rates. The median reporting period for pregnancy outcomes was 1 year; this ranged from 3 months to 10 years. Just three clinics reported up to date clinical pregnancy rates (covering the end of 2014). Twenty clinics (47% of those reporting clinical pregnancy) reported clinical pregnancy rates for multiple time periods, giving some indication of trends in performance.

Figure 4 (next page): Distribution of clinical outcome measures reported on medically assisted reproduction clinic websites. The denominator used is displayed for each numerator.

Clinic	Biochemical Pregnancy						Clinical Pregnancy															Continuing Pregnancy		Live birth								Cumulative Live birth
	Unspecified	Cycle started	Egg recovery	Frozen cycle	Insemination	Transfer	Unspecified	Day 5 transfer	Cycle started	Egg recovery	Embryo Transferred	Frozen cycle started	Insemination	Transfer	Cycle (unclear)	1st cycle (unclear)	Treatment (unclear)	Cumulative (unclear)	Course of inseminations	Egg recovery (cumulative)	3 cycles (cumulative)	Cycle started	Frozen cycle started	Unspecified	Day 5 transfer	Cycle started	Embryo Transferred	Frozen cycle started	Insemination	Transfer	Cycle (unclear)	Patient (unclear)
1						■								■	■																■	
2							■																									
3							■																									
4									■					■																		
5									■					■												■				■		
6							■									■																
7							■																									
8																										■						
9									■					■										■		■			■	■		
10									■																	■				■		■
11					■	■								■																		
12							■							■																		
13																										■	■	■				
14							■		■		■															■	■	■				
15														■																■		
16							■																									
17														■	■											■			■	■		
18	■						■			■				■												■		■				
19									■																							
20							■		■					■												■		■		■		
21										■		■	■	■																		
22																										■						
23						■								■										■		■						
24																																
25									■	■				■																		
26							■		■			■	■	■										■		■						
27		■	■			■			■	■		■	■	■																		
28	■	■							■					■												■				■		
29																																

Clinic	Biochemical Pregnancy						Clinical Pregnancy															Continuing Pregnancy		Live birth								Cumulative Live birth
	Unspecified	Cycle started	Egg recovery	Frozen cycle	Insemination	Transfer	Unspecified	Day 5 transfer	Cycle started	Egg recovery	Embryo Transferred	Frozen cycle started	Insemination	Transfer	Cycle (unclear)	1st cycle (unclear)	Treatment (unclear)	Cumulative (unclear)	Course of inseminations	Egg recovery (cumulative)	3 cycles (cumulative)	Cycle started	Frozen cycle started	Unspecified	Day 5 transfer	Cycle started	Embryo Transferred	Frozen cycle started	Insemination	Transfer	Cycle (unclear)	Patient (unclear)
30									■					■												■						
31									■					■																		
32																																
33									■		■			■												■	■					
34							■							■																■		
35																	■				■											
36														■																		
37																										■						
38														■																■		
39							■		■	■				■	■			■	■	■												
40								■	■																■	■			■	■	■	
41																										■		■				
42							■			■				■																		
43									■			■	■													■		■				
44																						■	■			■		■				
45									■																							
46		■							■																	■						
47														■																		
48							■			■				■																		
49		■	■	■	■	■								■																		
50	■																									■						
51														■																■		
52	■					■								■																		
53														■																		

Figure 5: Continuation of Figure 4

Live birth outcomes

Just over half (51%) of the clinics reported live birth rates, with 9 different live birth outcomes identified (Table 2). In contrast to pregnancy outcomes, it was most common to report live birth per cycle started (42% of clinics) as opposed to per transfer procedure (21%), perhaps reflecting the use of live birth as a patient-orientated outcome. A small number (6%) reported live birth per embryo transferred, although it could not be ascertained whether this was genuinely what was being reported or if this phrase had been used erroneously. A small number of websites (6%) reported live birth rate without defining the denominator. Just one website reported live birth rates cumulatively. These were reported 'per patient', although it was unclear at what point patients' data were censored. This website also reported the average number of cycles for those who achieved live birth (1.6), although this does not convey information about the expected number of cycles required to a patient faced with the decision of whether or not to commence IVF.

Only one clinic reported live birth per cycle started in such a way that patient age, sample size, included treatments, inclusion of fresh and/or frozen cycles, inclusion of donor cycles and reporting period were all clear. Nine (17%) clinics reported live birth per cycle started with each of age, sample size and period. Live birth rates were reported for a median time period of 1 year. However, this ranged from 3 months to 14 years. It is unclear how valid live birth rates can be reported for such short periods (the 3 month rates come from one clinic, the only one reporting live birth for a period of less than one year). Just three clinics reported live birth rates that were up to date (results from 2013 would have been available at the time of this review), although one of these stated that the results covered the whole of 2014, which is not possible given the follow up period required to establish live birth. Ten clinics (37% of clinics reporting live birth rates) reported live birth rates for multiple calendar periods, providing evidence of trends in performance.

Numerator	Denominator	No (%) of clinics reporting item
Biochemical pregnancy		10 (19% of clinics)
	unspecified denominator	4 (8)
	per cycle started	4 (8)
	per egg recovery	2 (4)
	per frozen cycle	1 (2)
	per insemination (IUI)	2 (4)
	per transfer procedure	6 (11)
Clinical pregnancy		43 (81% of clinics)
	unspecified denominator	14 (26)
	per day 5 transfer	1 (2)
	per cycle started	19 (36)
	per egg recovery	7 (13)
	per embryo transferred	2 (4)
	per frozen cycle started	4 (8)
	per insemination (IUI)	4 (8)
	per transfer procedure	29 (55)
	per cycle (ambiguous)	3 (6)
	per first cycle (ambiguous)	1 (2)
	per treatment (ambiguous)	1 (2)
	unspecified denominator (cumulative)	1 (2)
Clinical pregnancy (*cont*)		
	per course of inseminations (IUI)	1 (2)
	per egg collection (cumulative)	1 (2)
	per three cycles (cumulative)	1 (2)
Pregnancy (unspecified)		11 (21% of clinics)
	per patient (cumulative)	1 (2)
	per three cycles (cumulative)	1 (2)
	unspecified denominator	8 (15)
	per day 5 transfer	1 (2)
	per cycle started	3 (6)
	per frozen cycle started	1 (2)
	per insemination (IUI)	1 (2)
	per transfer procedure	2 (4)
	per cycle (ambiguous)	2 (4)
Singleton pregnancy		1 (2% of clinics)
	Unspecified denominator	1 (2)
Continuing pregnancy		1 (2% of clinics)
	per cycle started	1 (2)
	per frozen cycle started	1 (2)

Table 1: Reported pregnancy outcomes. Number (%) of clinics reporting each outcome.

Numerator	Denominator	No (%) of clinics reporting item.
Live birth		27 (51% of clinics)
	Unspecified denominator	3 (6)
	per day 5 transfer	1 (2)
	per cycle started	22 (42)
	per embryo transferred	3 (6)
	per frozen cycle started	7 (13)
	per insemination (IUI)	2 (4)
	per transfer procedure	11 (21)
	per cycle (ambiguous)	2 (4)
Cumulative live birth		1 (2% of clinics)
	per patient	1 (2)

Table 2: Reported live birth outcomes. Number (%) of clinics reporting each outcome.

Multiple Births

Eleven (21%) clinics reported information on multiple birth or pregnancies. Six (11%) clinics reported multiple birth rates. These were reported per live birth (two clinics), per cycle (one clinic) or without specifying the denominator (three clinics). Eight (15%) clinics reported multiple clinical pregnancy or multiple pregnancy rates. The denominator was either unspecified (four clinics) or per pregnancy (four clinics).

Pre-clinical outcomes

Just two clinics reported pre-clinical outcomes. Blastocyst achievement (with no denominator), implantation (no denominator) and transfer achieved per frozen cycle each appeared on one site.

Adverse events

Only one clinic reported adverse outcomes. Ectopic pregnancy and miscarriage were reported, although denominators were not specified.

3.4.3. Reporting of contextual information

Of the 53 clinics reporting outcomes, 14 (26%) presented (at least some) outcomes without specifying the age of the patients, 38 (72%) presented outcomes without specifying the treatments, 38 (72%) presented outcomes without specifying the sample size and 12 (23%) presented outcomes without specifying the period these related to.

Forty-eight (91%) presented outcomes for which it was unclear whether or not donor gametes were used. Forty-two (84%) presented outcomes for non-IUI treatments where it was unclear whether included cycles were fresh, frozen, or both fresh and frozen. Fifty (94%) presented outcomes without specifying how many patents did not complete the treatment.

Inclusion/ exclusion criteria were not consistently reported. Criteria relating to BMI could not be found for 64 (82%) of the websites, to age for 67 (85%) of the websites, to previous attempts for any website, or to smoking status for 94% of websites. 63 (80%) sites did not appear to provide criteria relating to any of these characteristics.

3.4.4. Comparison of NHS and private clinics

A higher proportion of NHS clinics compared with private centres reported age (89% vs 66%), sample size (50% vs 17%), use of donor gametes (17% vs 6%), use of fresh or frozen embryos (24% vs 12%, excluding IUI treatments) and number of abandoned treatments (17% vs 0%) for all outcomes. The proportion of NHS (28%) and private (29%) centres specifying the treatments involved for all reported results was similar. More private clinics (80%) than NHS clinics (72%) reported the date range for all outcomes.

3.5 Discussion

The present review confirms inconsistency in clarity and coverage when advertising clinic success rates with only one meeting HFEA's own standards. In addition to selecting from a number of numerators and denominators, clinics may also report results for different combinations of treatments, fresh and frozen cycles, donor and non-donor cycles and for different calendar periods. The large number of numerators and denominators in use constitutes an obstacle to consumer-friendly reporting, as patients may struggle to understand subtle differences in outcome definitions and may be misled into making comparisons between centres on the basis of incommensurable results (Chetkowski, 2014). Allowance of open reporting without binding guidelines carries a high risk of selective reporting; there is scope for clinics to construct more favourable outcomes using the variety of building blocks available. These points were highlighted by direct comparisons with other clinics using a `league table' presentation on 9 (11%) of the 79 websites. League tables are known to be problematic due to differences in patient characteristics and imprecision in the results used to create them (Marshall, et al., 1998). In addition to choices relating to outcome definition, league tables additionally allow clinics to select which other centres to include. These tables were invariably constructed so that the comparison was favourable to the reporting clinic. In one case, two websites used the outcome `live birth per cycle started' as the basis for a comparative table. Despite displaying results for overlapping (but not identical) time periods, one table indicated a considerable advantage of the reporting clinic over its competitor, while the other indicated that the performance of both clinics was comparable. The results used in both tables could not be called inaccurate.

The review raises concerns relating to clarity of reported results, with implications for patient usability. Current reporting trends are to present results in such a way so that the included treatments and inclusion or exclusion of frozen or donor gametes are often unclear. Given the multiplicity of relevant factors, a plausible rationale for these practices is to maintain simplicity. Complexity does represent a concern, as stakeholders may have difficulty interpreting conditional risk presented in the form of frequencies and percentages (eg: Bramwell, et al., 2006). However, by obscuring the particular patients and treatments for which results are presented, omission of such relevant information may in fact serve to obfuscate what is being reported. It was also common to report outcomes without sample sizes and without indicating the number of cycle cancellations or otherwise incomplete treatments, with implications for understanding the precision and the prognostic relevance of the results, respectively.

An emphasis on pregnancy was evident, with pregnancy outcomes representing the most common way to report success. The most common denominator used when reporting pregnancies was per transfer procedure. Considerably fewer clinics reported live birth rates. In contrast to pregnancy results, it was most common for these to be reported per cycle started. It has been argued that live birth is the most relevant measure of success of MAR to patients owing to the fact that this is the goal of any initiated treatment (Heijnen, et al., 2004, Malizia, et al., 2009, Min, et al., 2004, Moragianni and Penzias, 2010, Schieve and Reynolds, 2004, Tiitinen, et al., 2004) and that it is more informative to include all patients commencing treatment by counting events per cycle started (Heijnen, et al., 2004, Min, et al., 2004). Given that patients often undergo multiple attempts as part of their treatment, a case may be made for success rates to be presented cumulatively across some set time period or number of cycles (Gnoth, et al., 2011, Heijnen, et al., 2004, Luke, et al., 2012, Maheshwari, et al., 2015, Olivius, et al., 2002, Pelinck, et al., 2007, Soullier, et al., 2008, Stern, et al., 2010, Stewart, et al., 2011, Sundstrom and Saldeen, 2009, Witsenburg, et al., 2005). We found very few instances of this in the present study. This may be due to the practical challenges of calculating these cumulative rates and the need for a lengthy delay in reporting. HFEA have indicated that they will include cumulative live birth rates on their own website in future however. It is important to recognise that different outcomes may be suitable for different purposes, so that no single measure of success can be recommended. One proposal is that, whereas live birth per cycle started or per course of treatment may hold greater prognostic value, ongoing pregnancy may be more relevant for clinic performance evaluation (Griesinger, et al., 2004). A clear concern when deciding upon an appropriate performance measure is the impact that this may have on clinic behaviour. Clinics compete for patients, who are encouraged to consider performance when choosing a clinic (Johnson, et al., 2007). There is therefore an incentive to potentially modify the treatment delivered in order to optimise a particular performance indicator. This sort of gaming can lead to perverse behaviour which might not guarantee the best outcomes from a patient perspective (Bird, et al., 2005). This could manifest, for example, by clinics imposing tougher selection criteria, which we found to be sparsely reported (Sharif and Afnan, 2003). Without clearly presented selection policies, it is impossible to understand how much of a clinic's performance to attribute to treatment effectiveness and how much to the reproductive competence of their patients. We acknowledge that some centres may not have strict selection criteria, instead offering treatment to anyone who is able to pay. Nevertheless, it would be useful if these clinics reported that their results were based on relatively unselected cohorts. The desire to manipulate the behaviour of clinics to the advantage of patients motivates the proposal of live birth per embryo transferred as a measure of success, in order to encourage the transfer of fewer embryos at each attempt and to thereby reduce the incidence of multiple births (Abdalla, et al., 2010). On these grounds, HFEA plan to make live birth per embryo transferred the headline figure on their own website following their Information for Quality consultation (Human Fertilisation and

Embryology Authority, 2014). However, such a proposal introduces further complication as multiple embryos are not statistically independent.

Policies to reduce twin rates are ubiquitous outside the United States, and numbers of multiple births represent an important measure of clinic performance. Despite this, only 11 sites reported on multiple birth or pregnancy rates. Only one site reported on other adverse events. In the US, omission of information relating to side-effects has been noted as a characteristic of direct-to-consumer advertising of prescription drugs, with a substantial proportion of regulatory letters sent to manufacturers by the Food and Drug Administration (FDA) citing advertisements for minimisation of risks (Donohue, et al., 2007). It has been suggested that spending on direct to consumer advertising in the US increased drastically following changes to FDA regulations in 1997 that allowed manufacturers to advertise products without explicitly listing side-effects (Iizuka, 2004) although there is some evidence that the trend for increased spending actually preceded these changes (Ventola, 2011). In the present study, reporting of cancellations and abandoned treatments was also scanty, so that the actual chances of success for patients starting treatment could often not be discerned.

Our findings add to a body of literature highlighting the difficulty of reporting MAR outcomes in a consumer-friendly way. A 2007 review assessed US clinic websites according to the American Society for Reproductive Medicine/ Society for Assisted Reproductive Technology guidelines, and found generally low compliance (Abusief, et al., 2007). An earlier assessment of US clinic websites suggested generally low quality according to a scoring system based on American Medical Association internet health information guidelines (Huang, et al., 2005) although the methodology of the study has been queried given the status of these websites as advertisements (Epstein and Rosenberg, 2005, Jain and Barbieri, 2005). In the UK, a 2008 review of UK websites providing information on infertility found the quality of information to be variable, with particular concerns about accuracy (Marriott, et al., 2008). Quality control of data is essential for reliable performance monitoring (Bird, et al., 2005). At present, there is no way to guarantee the quality or accuracy of data presented on clinic websites.

The present study would appear to represent the first review of outcome reporting by UK MAR clinic websites. Strengths of the study include the extraction of item-level data, allowing the variety of outcomes in use by UK clinics to be presented. Limitations of the study should be noted. In particular, this review was cross-sectional, meaning that we are unable to comment on reporting trends over time. We have also not considered alternative ways in which clinics use the internet to communicate results to patients, such as social media. Our comparison of NHS and private clinics is also tentative; we used the presence or absence of the NHS logo on the front page of the site to distinguish NHS from private centres, with one exception (a private clinic where the logo was clearly used to illustrate an existing NHS contract). This method is obviously imperfect, and while we believe that we managed to correctly categorise clinics, it is possible that some misclassification occurred. With these limitations in mind, we conclude that self-regulation does not appear to guarantee clear, patient-friendly reporting of outcomes.

Our intention is not accusatory; the matter of how to report MAR outcomes is complex and we expect that many clinics present their success rates in good faith. There are clear parallels to ongoing discussions about the presentation of online information in other areas, such as cosmetic procedures (eg: Light, et al., 2014) or alternative medicine (eg: Beutel and Cardone, 2014). There is a tension between ‘open reporting’ in the interests of transparency and ‘direct to consumer advertising’, particularly for private providers. One method to address this would be binding guidance for consistent content in reporting results. Another would be an outright ban on direct advertising of MAR.

ACKNOWLEDGEMENT

We would like to thanks Daniel Brison for helpful discussions regarding this study and to James Duffy for his comments on an earlier draft of this manuscript, and specifically for suggesting the display presented as Figure 4. We would also like to thank Mika Gissler for suggesting the analyses investigating whether the most recent results were presented and whether trends were reported during the peer review of this manuscript. These had not occurred to us.

LICENCE STATEMENT

ETHICS STATEMENT

Ethical approval was not required as the study consisted of the analysis of publically available summary-level data.

DETAILS OF FUNDING

JW is funded by a Doctoral Research Fellowship from the National Institute for Health Research, supervised by AV and SR (DRF-2014-07-050). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

DECLARATION OF COMPETING INTERESTS

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: JW is funded by a Doctoral Research Fellowship from the National Institute for Health Research, supervised by AV and SR; AV and JW are statistical editors of the Cochrane Gynaecology and Fertility Group; no other relationships or activities that could appear to have influenced the submitted work.

TRANSPARENCY STATEMENT

JW affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

DETAILS OF THE ROLE OF THE STUDY SPONSORS

JW is funded by a Doctoral Research Fellowship granted by the NIHR. The NIHR approved a research plan including the present study. The NIHR did not contribute to the design or analysis of the study, nor were they involved in drafting or approving the manuscript.

STATEMENT OF INDPENDENCE OF RESEARCHERS FROM FUNDERS

This report is independent research arising in part from a Doctoral Research Fellowship supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

DATA SHARING STATEMENT

Copies of the datasets compiled and analysed for the present study may be acquired by contacting the corresponding author, and have been made available on the Mendeley Data repository.

3.6 References for Chapter 3.

Abdalla HI, Bhattacharya S, Khalaf Y. Is meaningful reporting of national IVF outcome data possible? *Hum Reprod* 2010;25: 9-13.

Abusief ME, Hornstein MD, Jain T. Assessment of United States fertility clinic websites according to the American Society for Reproductive Medicine (ASRM)/Society for Assisted Reproductive Technology (SART) guidelines. *Fertil Steril* 2007;87: 88-92.

Beutel BG, Cardone DA. KINESIOLOGY TAPING AND THE WORLD WIDE WEB: A QUALITY AND CONTENT ANALYSIS OF INTERNET-BASED INFORMATION. *Int J Sports Phys Ther* 2014;9: 665.

Bird SM, David C, Farewell VT, Harvey G, Tim H, Peter C. Performance indicators: good, bad, and ugly. *J R Stat Soc Ser A* 2005;168: 1-27.

Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. *Bmj* 2006;333: 284.

Chetkowski RJ. Consumer-friendly reporting of in vitro fertilization outcomes. *Fertil Steril* 2014;101: e7.

Donohue JM, Cevasco M, Rosenthal MB. A decade of direct-to-consumer advertising of prescription drugs. *N Engl J Med* 2007;357: 673-681.

Epstein YM, Rosenberg HS. Assessing infertility information on the Internet: Challenges and possible solutions. *Fertil Steril* 2005;83: 553-555.

Garrido N, Bellver J, Remohi J, Simon C, Pellicer A. Cumulative live-birth rates per total number of embryos needed to reach newborn in consecutive in vitro fertilization (IVF) cycles: a new approach to measuring the likelihood of IVF success. *Fertil Steril* 2011;96: 40-46.

Germond M, Urner F, Chanson A, Primi M-P, Wirthner D, Senn A. What is the most relevant standard of success in assisted reproduction? The cumulated singleton/twin delivery rates per oocyte pick-up: the CUSIDERA and CUTWIDERA. *Hum Reprod* 2004;19: 2442-2444.

Gnoth C, Maxrath B, Skonieczny T, Friol K, Godehardt E, Tigges J. Final ART success rates: a 10 years survey. *Hum Reprod* 2011;26: 2239-2246.

Griesinger G, Dafopoulos K, Schultze-Mosgau A, Felberbaum R, Diedrich K. What is the most relevant standard of success in assisted reproduction? Is BESST (birth emphasizing a successful singleton at term) truly the best? *Hum Reprod* 2004;19: 1239-1241.

Haagen EC, Tuil W, Hendriks J, de Bruijn RPJ, Braat DDM, Kremer JAM. Current Internet use and preferences of IVF and ICSI patients. *Hum Reprod* 2003;18: 2073-2078.

Heijnen E, Macklon NS, Fauser B. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod* 2004;19: 1936-1938.

Huang JY, Discepola F, Al-Fozan H, Tulandi T. Quality of fertility clinic websites. *Fertil Steril* 2005;83: 538-544.

Human Fertilisation and Embryology Authority. Choose a Fertility Clinic. 2009.

Human Fertilisation and Embryology Authority. Information for Quality Consultation. 2014.

Iizuka T. What explains the use of direct-to-consumer advertising of prescription drugs? *The Journal of Industrial Economics* 2004;52: 349-379.

Jain T, Barbieri RL. Website quality assessment: Mistaking apples for oranges. *Fertil Steril* 2005;83: 545-547.

Johnson A, El-Toukhy T, Sunkara S, Khairy M, Coomarasamy A, Ross C, Bora S, Khalaf Y, Braude P. Short communication: Validity of the in vitro fertilisation league tables: influence of patients' characteristics. *BJOG: An International Journal of Obstetrics & Gynaecology* 2007;114: 1569-1574.

Light A, Munro C, Breakey W, Critchley A. The Internet: What are our patients exposed to when considering breast reconstruction following mastectomy? *The Breast* 2014;23: 799-806.

Luke B, Brown MB, Wantman E, Lederman A, Gibbons W, Schattman GL, Lobo RA, Leach RE, Stern JE. Cumulative birth rates with linked assisted reproductive technology cycles. *N Engl J Med* 2012;366: 2483-2491.

Maheshwari A, McLernon D, Bhattacharya S. Cumulative live birth rate: time for a consensus? *Hum Reprod* 2015;30: 2703-2707.

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

Marriott JV, Stec P, El-Toukhy T, Khalaf Y, Braude P, Coomarasamy A. Infertility information on the World Wide Web: a cross-sectional survey of quality of infertility information on the internet in the UK. *Hum Reprod* 2008;23: 1520-1525.

Marshall EC, Sanderson C, Spiegelhalter DJ, McKee M. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth ratesCommentary: How robust are rankings? The implications of confidence intervals. *Bmj* 1998;316: 1701-1705.

Meldrum DR. Pregnancies and deliveries per fresh cycle are no longer adequate indicators of in vitro fertilization program quality: how should registries adapt? *Fertil Steril* 2013;100: 620.

Min JK, Breheny SA, MacLachlan V, Healy DL. What is the most relevant standard of success in assisted reproduction? The singleton, term gestation, live birth rate per cycle initiated: the BESST endpoint for assisted reproduction. *Hum Reprod* 2004;19: 3-7.

Moragianni VA, Penzias AS. Cumulative live-birth rates after assisted reproductive technology. *Curr Opin Obstet Gynecol* 2010;22: 189-192.

Olivius K, Friden B, Lundin K, Bergh C. Cumulative probability of live birth after three in vitro fertilization/intracytoplasmic sperm injection cycles. *Fertil Steril* 2002;77: 505-510.

Pelinck M, Vogel N, Arts E, Simons A, Heineman M, Hoek A. Cumulative pregnancy rates after a maximum of nine cycles of modified natural cycle IVF and analysis of patient drop-out: a cohort study. *Hum Reprod* 2007;22: 2463-2470.

Pinborg A, Loft A, Ziebe S, Andersen AN. What is the most relevant standard of success in assisted reproduction? Is there a single ‘parameter of excellence’? *Hum Reprod* 2004;19: 1052-1054.

Rawal N, Haddad N. Use of Internet in infertility patients. *The Internet Journal of Gynecology and Obstetrics* 2006;5.

Schieve LA, Reynolds MA. What is the most relevant standard of success in assisted reproduction? Challenges in measuring and reporting success rates for assisted reproductive technology treatments: What is optimal? *Hum Reprod* 2004;19: 778-782.

Sharif K, Afnan M. The IVF league tables: time for a reality check. *Hum Reprod* 2003;18: 483-485.

Soullier N, Bouyer J, Pouly JL, Guibert J, de La Rochebrochard E. Estimating the success of an in vitro fertilization programme using multiple imputation. *Hum Reprod* 2008;23: 187-192.

Stern JE, Brown MB, Luke B, Wantman E, Lederman A, Missmer SA, Hornstein MD. Calculating cumulative live-birth rates from linked cycles of assisted reproductive technology (ART): data from the Massachusetts SART CORS. *Fertil Steril* 2010;94: 1334-1340.

Stewart LM, Holman CAJ, Hart R, Finn J, Mai Q, Preen DB. How effective is in vitro fertilization, and how can it be improved? *Fertil Steril* 2011;95: 1677-1683.

Sundstrom P, Saldeen P. Cumulative delivery rate in an in vitro fertilization program with a single embryo transfer policy. *Acta Obstet Gynecol Scand* 2009;88: 700-706.

Talarczyk J, Hauke J, Poniewaz M, Serdynska-Szuster M, Pawelczyk L, Jedrzejczak P. Internet as a source of information about infertility among infertile patients. *Ginekologia Polska* 2012;83: 250-254.

Tiitinen A, Hydén-Granskog C, Gissler M. What is the most relevant standard of success in assisted reproduction? The value of cryopreservation on cumulative pregnancy rates per single oocyte retrieval should not be forgotten. *Hum Reprod* 2004;19: 2439-2441.

Ventola CL. Direct-to-consumer pharmaceutical advertising: therapeutic or toxic? *P&T* 2011;36: 669.

Witsenburg C, Dieben S, Van der Westerlaken L, Verburg H, Naaktgeboren N. Cumulative live birth rates in cohorts of patients treated with in vitro fertilization or intracytoplasmic sperm injection. *Fertil Steril* 2005;84: 99-107.

Zegers-Hochschild F, Adamson GD, de Mouzon J, Ishihara O, Mansour R, Nygren K, Sullivan E, Vanderpoel S. International Committee for Monitoring Assisted Reproductive

Technology (ICMART) and the World Health Organization (WHO) revised glossary of ART terminology, 2009∗. *Fertil Steril* 2009;92: 1520-1524.

Chapter 4. No common denominator: a review of outcome measures in IVF randomised controlled trials.

Journal article 2

Authors Jack Wilkinson, Stephen A Roberts, Marian Showell, Daniel R Brison, Andy Vail.

Status Published in Human Reproduction

Reference Wilkinson, J., Roberts, S.A., Showell, M., Brison, D.R., Vail, A. (2016). "No common denominator: a review of outcome measures in IVF RCTs." Hum Reprod 31(12): 2714-2722.

Contribution statement JW designed the study, and undertook the acquisition, analysis and interpretation of data, drafted the manuscript and gave final approval of the version to be published. MS designed and conducted the search, contributed to the design of the study, the interpretation of data, drafting and revision of the manuscript and gave final approval of the version to be published. All other authors contributed to the design of the study, the interpretation of data, drafting and revision of the manuscript and gave final approval of the version to be published. JW is acting as guarantor for the study.

Preamble As for our review of clinic websites, our primary motivation when conducting this review had been to inform method development. As for that review however, it became apparent that there were important methodological implications arising from reporting heterogeneity in RCTs. Again, we felt that it was important to highlight this issue.

Outputs and Impact of the research This work was presented in poster format at the 2016 conferences of The International Society for Clinical Biostatistics and The Royal Statistical Society. At the time of writing, JW has been invited to take part in the COMMIT (Core Outcome Measures for Infertility Trials and Priority Setting for Infertility) project, and it is anticipated that the dataset underlying this study will form the basis of this work.

4.1 **Abstract**

Study question

Which outcome measures are reported in randomised controlled trials (RCTs) for in vitro fertilisation (IVF)?

Summary answer

Many combinations of numerator and denominator are in use, and are often employed in a manner that compromises the validity of the study.

What is known already

The choice of numerator and denominator governs the meaning, relevance and statistical integrity of a study's results. RCTs only provide reliable evidence when outcomes are assessed in the cohort of randomised participants, rather than in the subgroup of patients who completed treatment.

Study design, size, duration

Review of outcome measures reported in 142 IVF RCTs published in 2013 or 2014.

Participants/materials, setting, methods

Trials were identified by searching the Cochrane Gynaecology and Fertility Specialised Register. Reported numerators and denominators were extracted. Where they were reported, we checked to see if live birth rates were calculated correctly using the entire randomised cohort or a later denominator.

Main results and the role of chance

Over 800 combinations of numerator and denominator were identified. No single outcome measure appeared in the majority of trials. Only 22 (43%) studies reporting live birth presented a calculation including all randomised participants or only excluding protocol violators. A variety of definitions were used for key clinical numerators.

Limitations, reasons for caution

Several of the included articles may have been secondary publications. Our categorisation scheme was essentially arbitrary, so the frequencies we present should be interpreted with this in mind. The analysis of live birth denominators was post-hoc.

Wider implications of the findings

There is massive diversity in numerator and denominator selection in IVF trials due to its multistage nature, and this causes methodological frailty in the evidence base. The twin spectres of outcome reporting bias and analysis of non-randomised comparisons do not appear to be widely recognised. Initiatives to standardise outcome reporting are welcome, although there is a need to recognise that early outcomes of treatment may be appropriate choices of primary outcome for early phase studies.

Study funding/ competing interests

JW is funded by a Doctoral Research Fellowship from the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. JW also declares that publishing research is beneficial to his career. JW and AV are statistical editors, and MS is Information Specialist, for the Cochrane Gynaecology and Fertility Group, although the views expressed here are not necessarily those of the group. DRB is funded by the NHS as Scientific Director of a clinical IVF service. The authors declare no other conflicts of interest.

KEY WORDS

In vitro fertilisation. Outcome measures. Assisted reproduction. Core outcomes. Live birth.

4.2 **Introduction**

Inconsistency and incompleteness of outcome reporting in infertility trials are barriers to understanding and improving treatments (Dapuzzo, et al., 2011, Legro, et al., 2014). In the absence of common standards of reporting, it may be difficult to compare the safety and effectiveness of competing interventions, or to synthesise the results of trials in meta-analysis (Blazeby, et al., 2012, Clarke and Williamson, 2016, Khan, 2014). The choice of outcome also has implications for both the relevance (Heijnen, et al., 2004, Legro, et al., 2014, Min, et al., 2004) and methodological validity (Griesinger, 2016, Vail and Gardener, 2003) of a trial's results.

Choosing an outcome for trials of in vitro fertilisation (IVF) is particularly complex, owing to the multistage nature of the treatment. Treatment comprises stimulation of the ovaries, retrieval and fertilisation of oocytes and the culture and transfer of some of the resulting embryos to the uterine cavity (Van Voorhis, 2007). Some of these embryos may implant, some of these may result in a clinical pregnancy, and some of these may result in a live birth. Those embryos not used for the initial transfer may be cryopreserved, so that they can later be thawed and transferred in a subsequent attempt. The response at each stage can be quantified: ovarian response by the number and maturity of oocytes; fertilisation by the number of zygotes, and subsequently the number and quality of embryos produced; the transfer procedure by the implantation of embryos; and the clinical outcome of treatment by clinical pregnancy and the birth of a child. Additionally, treatment may fail at each stage: stimulation may be cancelled due to poor or overresponse; fertilisation failure may occur; embryos may fail to develop, or post transfer fail to implant; and pregnancies may be lost before or subsequent to identification of a clinical pregnancy. One consequence of this for clinical trials of interventions designed to improve IVF is that numerous clinical and procedural events that occur during treatment can be reported. A second consequence is that these events may be reported in subgroups containing only those patients who reach a certain milestone, such as oocyte retrieval or embryo transfer. Further complexity arises due to the fact that IVF involves two or more individuals (for example a male and female partner), who may undertake multiple treatment cycles, and one or more additional individuals (babies) arising from successful treatment (Legro, et al.,2014). When selecting which outcomes to report in an IVF trial therefore, many numerators and denominators are available (Heijnen et al., 2004).

The importance of the choice of numerator is well recognised and has been enshrined in the IMPRINT (Improving the Reporting of Clinical Trials of Infertility Treatments) statement with a call for live birth to be reported in all infertility trials (Legro, et al., 2014), although alternatives, such as ongoing pregnancy, have been proposed on pragmatic grounds (Braakhekke, et al., 2014a). The appropriate choice of denominator is a more subtle issue. The optimal denominator for IVF evaluation has been widely discussed (Abdalla, et al., 2010, Garrido, et al., 2011, Germond, et al., 2004, Heijnen, et al., 2004), and is known to have implications for the interpretation of trials, where the exclusion of randomised participants may introduce bias to the estimated treatment effect (Montori and Guyatt, 2001, Vail and Gardener, 2003, Mastenbroek, et al., 2005, Mastenbroek and Repping, 2014). This could occur if, for example, participants are randomised at the start of ovarian stimulation, but the outcome is calculated only in those who undergo transfer.

We conducted a review of outcomes reported in IVF randomised controlled trials in 2013 and 2014. Our aims were to establish the full range of outcomes in use in IVF randomised controlled trials (RCTs) and to identify the ramifications for the evidence base.

4.3 Methods

4.3.1. Search strategy

MS performed a search of the Cochrane Gynaecology and Fertility Group PROCITE database on 22/06/15 using the search strategy contained in Appendix 1. This is a specialised register of RCTs updated weekly by searching databases, conference abstracts and journals. Further details of the database are provided in Appendix 2. Our initial search covered the period 2010 to 2014, although we subsequently narrowed our focus to the period 2013 to 2014 due to feasibility constraints. We screened the titles and abstracts of the identified articles and excluded those not meeting the eligibility criteria. We reviewed the full text of all articles not excluded during this initial screening phase and made further exclusions as appropriate.

Eligibility criteria

English-language publications of randomised controlled trials in peer-reviewed journals in the period 1st January 2013 to 31st December 2014 were considered eligible. Conference papers were excluded. We did not consider methodological quality to be relevant, as our concerns related to the outcomes reported in this literature and not in the estimation of treatment effects. To be eligible, a study had to have had participants undergoing IVF or intracytoplasmic sperm injection (ICSI) including a period of ovarian stimulation in at least one arm of the trial, or participants undergoing frozen embryo transfer in at least one arm of the trial, or partners of patients undergoing IVF or ICSI in at least one arm of the trial, or oocyte donors donating to an IVF programme. We included trials where surplus oocytes had been obtained as part of IVF or ICSI treatment and an intervention was applied to these oocytes even if there was no intention to subsequently transfer any of the resulting embryos. Finally, the publication had to report clinical or preclinical outcomes to be eligible (which would exclude, for example, purely economic evaluations of interventions).

4.3.2. **Data extraction**

Initially, we performed a small pilot extraction of 5 reports to inform the extraction process used in the full sample, including the variables to be extracted and the formatting of this information. We extracted information at both study-level and at the level of each reported outcome in a study. We defined an outcome as any post-randomisation variable presented separately for each arm in the study or as a comparison between study arms and recorded both the numerator and denominator used in the calculation. We did not record a reported outcome multiple times if it was presented for each of several subgroups, unless these were defined by excluding patients who did not reach a certain stage in the process. We also did not record outcomes multiple times where these corresponded to repeated measurements at several timepoints. At the study-level, we extracted details of the intervention and the stage in the treatment process at which the intervention was applied (pre-stimulation phase, stimulation phase, post-stimulation including culture and selection of embryos, transfer, frozen transfer or intervention targeted at the male partner, such as manipulation or selection of sperm prior to ICSI). Similarly, we extracted the stage of treatment at which randomisation took place. For each reported outcome, we extracted the numerator and denominator (for numerical variables, the denominator would be the divisor used in the calculation of a mean). Where pregnancy or live birth were reported, we extracted the corresponding definition used by the study authors. Data were extracted into two databases, one containing study-level information and another containing reported-outcome-level information. JW performed data extraction for all studies. SR and AV performed double extraction for a random sample of 10%, to check data quality and consistency of recording. Furthermore, we conducted extensive data validation and cleaning, including manually checking every entered item.

4.3.3. **Statistical analysis**

We summarised the characteristics of the sample and tabulated the numerators and denominators in use in 9 categories (live birth, pregnancy, stimulation response, transfer, fertilisation, multiple births or pregnancies, other preclinical outcomes, adverse events, postnatal). These categories are arbitrary and have been selected to facilitate the presentation of our results. We note here however that, since our analyses are descriptive and these categories are purely presentational, it would not affect our results were an outcome measure to be reported under one heading rather than another. Due to the large number of outcomes identified, we reported only those appearing in more than one study. We simplified the results by combining similar numerators and denominators. For example, we combined live birth with take home baby rate, and combined the denominators 'per patient with sufficient embryos' and 'per patient with sufficient blastocysts', where 'sufficiency' could be defined on the basis of quantity or quality of embryos (or both). For this primary analysis, we did not distinguish between subtly different definitions of outcomes (for example, clinical pregnancy may have been defined as foetal heartbeat on ultrasound at different timepoints in different studies). However, at the suggestion of an anonymous peer reviewer, we also present the definitions used by trial authors for pregnancy and live birth outcomes. In order to investigate the methodological implications of denominator selection, we conducted post-hoc analysis in the subgroup of studies reporting live birth. We recorded whether the denominator used coincided with the cohort of randomised participants (ignoring exclusions due to protocol violations) and if not, the nature and extent of the exclusion. We did not perform statistical inference, because we have attempted to summarise all trials within the time period and it isn't clear that inference would be meaningful.

4.3.4. Sample size

The decision to include all studies in the period 01/01/13 to 31/12/14 was made primarily on pragmatic grounds, on the basis that this would be sufficient to assess current practices in outcome reporting while proving to be feasible. A post-hoc calculation can be made however. A sample of size 142 yields a 76% probability of observing a relatively rare outcome (appearing in 1 out of every 100 studies) at least once.

4.3.5. Ethical approval

Ethical approval was not required as the study involved only the review of published research.

4.4 Results

4.4.1. Results of the search

Figure 6 shows the results of the search and screening process. The search identified 640 references published between 2013 and 2014. Following title and abstract screening, 488 references were discarded without further assessment. The remaining 152 articles were assessed further by reviewing the full texts and a further 10 were excluded for the reasons shown in Figure 6. 142 RCTs were included in the analysis. Agreement between raters was almost universal, with one reviewer erroneously extracting one additional outcome from one study due to misreading the text.

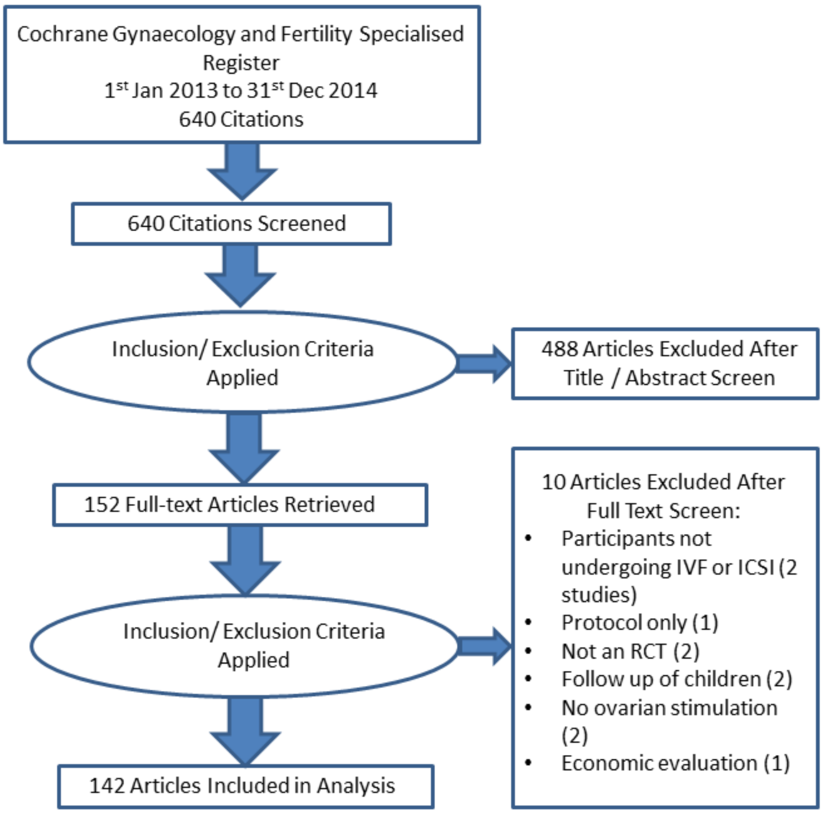

Cochrane Gynaecology and Fertility Specialised Register
1st Jan 2013 to 31st Dec 2014
640 Citations
640 Citations Screened
Inclusion/ Exclusion Criteria Applied
488 Articles Excluded After Title / Abstract Screen
152 Full-text Articles Retrieved
Inclusion/ Exclusion Criteria Applied
10 Articles Excluded After Full Text Screen:
• Participants not undergoing IVF or ICSI (2 studies)
• Protocol only (1)
• Not an RCT (2)
• Follow up of children (2)
• No ovarian stimulation (2)
• Economic evaluation (1)
142 Articles Included in Analysis

Figure 6: PRISMA Diagram showing flow of studies in the review.

4.4.2. Stage of intervention and randomisation

Interventions were delivered prior to ovarian stimulation in 20 (14%) articles, during stimulation in 51 (36%), post stimulation or during culture of embryos in 31 (22%), post culture but preceding transfer of embryos in 19 (13%) and following the transfer procedure in 3 (2%). Five (4%) were trials of interventions targeted at male partners and 13 (9%) featured interventions designed to improve outcomes after the vitrification and warming of oocytes or embryos. Randomisation occurred prior to stimulation in 62 (44%) articles, during stimulation in 17 (12%), post stimulation or during culture in 27 (19%) and post culture but prior to transfer in 23 (16%). The point of randomisation was unclear in 13 (9%) articles.

4.4.3. Reported outcomes

After combining similar items, 361 numerators and 87 denominators were discerned. 815 distinct combinations of numerator and denominator were identified. 203 combinations appeared in more than one study (612 did not). The median (interquartile range) of distinct outcomes reported in a study was 11 (7 to 16), with a range of 1 to 36.

Live birth outcomes

Fifty-two (37%) articles reported the numerators live birth event or take home baby in total, with 14 combinations of numerator and denominator. Figure 7 and S Table 1 show combinations of live birth numerators and denominators appearing in more than one study. It was most common to report these per transfer (15% of studies). Only 8 (6%) studies reported live birth per cycle started. It was not common (5%) for studies to report live birth in a cumulative fashion, across multiple fresh and frozen transfer cycles. No study reported cumulative live birth following multiple egg collections. Four (3%) reported cumulative live birth per cycle started and 2 (1%) reported time to pregnancy leading to live birth, where time was measured across multiple treatment cycles. Four (3%) of studies reported preterm birth event with three of these reporting preterm birth per baby.

Of the 52 studies reporting live birth rates, 22 (42%) used the point of randomisation as the denominator in the calculation. One study acknowledged that the calculation was not based on a randomised comparison and was therefore ‘descriptive’. In six (12%) studies, the denominator could not be discerned. The remaining 23 (44%) did not use the randomised cohort as the denominator. In seventeen (33%) studies, the denominator included only those undergoing transfer (15 studies) or oocyte retrieval (two studies) rather than the randomised participant. In these 17 studies, a median (IQR) of 8% (4 to 14%) of participants were excluded, with a range of (2 to 38%). Seven (13%) studies made a unit of analysis error when calculating live birth rates, with six calculating live birth rates per embryo transferred. In one trial each woman’s oocytes were randomly split between intervention arms, and live birth per transfer was calculated in the subset of procedures where all embryos transferred had originated from one of the arms.

Pregnancy outcomes

Table 3 shows pregnancy outcomes appearing in more than one study. 46 (32%) reported biochemical pregnancy, with 13 different denominators. It was most common (16%) to report these per transfer procedure. Clinical pregnancy rates (with varying definitions) were reported in most (67%) studies, with 19 different denominators. Again, it was most common to report these per transfer procedure (31%) although the denominator ‘per cycle started’ was also reasonably prevalent (17%). Thirty-nine (27%) studies reported ongoing pregnancy using 16 different denominators, with 15% reporting ongoing pregnancy per transfer procedure. Only 5% reported ongoing pregnancy per cycle started. Very few studies reported clinical pregnancy (1%) or ongoing pregnancy (2%) in a cumulative fashion. Just under half (43%) reported miscarriages in addition to 6% reporting pregnancies that did not progress beyond the biochemical stage. Nineteen (13%) reported miscarriages per clinical pregnancy and 11 (8%) reported these per biochemical pregnancy.

Figure 7 (opposite): Reported live birth outcomes in IVF RCTs in 2013-2014 by stage of intervention (A to F). Each row corresponds to a single study. Only studies reporting live birth outcome measures appearing in more than one study are shown. Blue triangles (▲) indicate that the study authors used a denominator that coincided with the point of randomisation in the trial. Red circles (●) indicate that the study authors did not use the point of randomisation as the denominator, but instead included only patients who reached a certain stage of treatment when calculating live birth rates, potentially undermining the random allocation in the study. Black squircles (■) indicate that it is unclear whether or not the denominator coincided with the point of randomisation.

[1] Authors presented this as a descriptive result

Study	Live birth event or take home baby							Cumulative live birth	
Denominator	Cycle started (or earlier)	Patient achieving trigger	Oocyte retrieval	Patient w/ sufficient embryos	Transfer procedure	Embryo transferred	Unclear denominator	star Course of ted treatment	Time to pregnancy resulting in birth
A: Male intervention (2 of 5 articles report live birth)									
1					●				
2					●				
B: Pre-stimulation (9 of 20 articles report live birth)									
3					●				
4	▲								
5								▲	▲
6					●				
7					●				
8	▲								
9	▲								
10	▲				●				
11							■		
C: Stimulation phase (12 of 51 articles report live birth)									
12	▲							▲	
13	▲				●			▲	▲
14	▲								
15					●				
16		▲			●				
17					●				
18			●						
19	▲								
20			▲			●			
21									
22		▲							
23			●						
D: Post-stimulation or during culture (13 of 31 articles report live birth)									
24					●				
25					●				
26					●[1]				
27			●						
28						●			
29							■		
30					●				
31							■		
32							■		
33					●				
34			●		●				
35			▲						
36					●				
E: Post culture but prior to transfer of embryos (10 of 19 articles report live birth)									
37					●		■		
38					●				
39				▲		●			
40						●			
41						●			
42								▲	
43				▲					
44							■		
45				▲		●			
46				▲					
F: Post transfer (1 of 3 articles report live birth)									
47				▲					

Stimulation outcomes

S Table 2 and S Table 3 show outcomes relating to stimulation response. Number of oocytes (46%), of mature oocytes (23%), total gonadotropin dose (27%) and stimulation duration (26%) were all commonly reported, each with a variety of denominators. Perhaps unsurprisingly, stimulation outcomes were more frequently reported per cycle started compared to pregnancy and live birth events; 28 (20%) reported number of oocytes, 17 (12%) reported number of mature oocytes, 21 (15%) reported gonadotropin dose and 19 (13%) reported stimulation duration per cycle started. However, some studies did report stimulation outcomes in the subset of patients reaching later stages in the process (S Table 3). Eighteen (13%) studies reported cycle cancellation, 13 (9%) per cycle started.

Fertilisation outcomes

S Table 4 and S Table 5 show fertilisation outcomes. Fertilisation (37%), the attainment of good quality embryos as a binary variable (15%), the number of embryos (19%), of good quality embryos (12%) and of frozen embryos (14%) were all frequently reported, each with a variety of denominators (S Table 4, S Table 5). Other than cleavage (11%), no other numerator was reported in more than 8% of studies

Numerator	Denominator	No (%) of studies	Numerator	Denominator	No (%) of studies
Biochemical pregnancy		46 (32% of studies)	*Ongoing pregnancy*		39 (27%)
	per cycle started (or earlier)	12 (8%)		per cycle started (or earlier)	7 (5%)
	per transfer	23 (16%)		per oocyte retrieval	5 (4%)
	per patient achieving trigger	2(1%)		per patient with sufficient embryos	5 (4%)
	per oocyte retrieval	2(1%)		per transfer	21 (15%)
	par patient w/ sufficient embryos	5(4%)		per clinical pregnancy	3 (2%)
	unclear denominator	2(1%)	*Pregnancy (unclear)*		9 (6%)
Biochemical pregnancy only		9 (6% of studies)		per cycle started (or earlier)	2 (1%)
	per transfer	2 (1%)		per transfer	4 (3%)
	per transfer of embryos from one intervention arm only	2 (1%)	*Cumulative clinical pregnancy*		2 (1%)
	per chemical pregnancy	2 (1%)		per course of treatment started	2 (1%)
	unclear	2(1%)	*Cumulative ongoing pregnancy*		3 (2%)
Clinical pregnancy		95 (67% of studies)		per course of treatment started	2 (1%)
	per cycle started (or earlier)	24 (17%)	*Miscarriage*		61 (43%)
	per trigger	4 (3%)		per chemical pregnancy	11 (8%)
	per oocyte retrieval	11 (8%)		per clinical pregnancy	19 (13%)
	per patient w/ sufficient embryos	6 (4%)		per cycle started (or earlier)	3 (2%)
	per transfer	44 (31%)		per oocyte retrieval	3 (2%)
	per transfer of embryos from one intervention arm only	3 (2%)		per transfer	9 (6%)
	unclear	7 (5%)		per transfer of embryos from one intervention arm only	2 (1%)
	per clinical pregnancy	19 (13%)		unclear	9 (6%)

Table 3: Pregnancy outcomes reported by more than one study. Frequency (%) of studies reporting each outcome.

Transfer outcomes

S Table 6 and S Table 7 display outcomes relating to the transfer procedure. Number of embryos transferred (52%) and implantation (52%) were the most commonly reported numerators in the review. The denominator used with implantation was often unclear (38%) but was otherwise generally reported per embryo transferred (30%) rather than as a patient-level outcome. Number of embryos transferred was most commonly reported per transfer procedure (17%). Other transfer outcomes appeared in relatively small numbers of studies; the next most recurrent was achievement of transfer (8%), reported per cycle started (4%) or per oocyte retrieval (2%).

Multiple pregnancies and births

Relatively few studies reported multiple pregnancies or births or pregnancies (S Table 8). 17% reported the numerator multiple pregnancy and 4% reported multiple birth rates. One study reported multiple pregnancy per cycle started, the only instance of an outcome in this category being reported with this denominator. Where multiple pregnancy was reported, it was not uncommon for it to be presented per clinical pregnancy (5%). Multiple birth was only reported per live birth event (3%) or per transfer (1%).

Other adverse events

The most commonly reported adverse event was ovarian hyperstimulation syndrome (OHSS) of unspecified severity (17%), with several studies specifying the severity as mild (3%), moderate (4%) or severe (4%) (S Table 9). Ectopic pregnancy rates were explicitly reported in 13% of studies and general adverse events were described in 6%.

Postnatal outcomes

Small numbers of studies reported postnatal outcomes, most commonly birthweight (6%), congenital abnormalities (4%) and gestational age (2%) (S Table 10). These were most frequently reported per baby.

Other procedural outcomes

Other procedural measurements were reasonably prevalent, such as estradiol levels (32%), endometrial thickness (25%) or progesterone levels (12%) (S Table 11). These outcomes were generally reported using denominators including patients in the earlier stages of treatment (eg: per cycle started or per oocyte retrieval).

4.4.4. Definitions of pregnancy and live birth used in the studies

Note that for these analyses, we have included the definitions used when variants of these outcome measures were reported, for example giving the definition of live birth used when cumulative live birth was reported. Accordingly, the totals for these analyses do not match those in the analyses described above.

Live birth

Table 4 shows the definitions provided by authors reporting live birth. It was most common (27 studies, 51%) for no definition to be given, followed by 19 (36%) defining this as a count of live birth events/deliveries. Other definitions, such as counts of individual babies, were sparse.

Clinical Pregnancy

S Table 12 shows the definitions of clinical pregnancy. This was not defined in around one fifth (21, 21%) of studies reporting clinical pregnancy. A variety of subtly different definitions were used, with the vast majority comprising some combination of ultrasound confirmation of gestational sacs and foetal heartbeat at different timepoints.

Definition of live birth	Frequency (%) of studies
Birth of >=1 neonate 28 weeks or later	1 (2)
Individual baby born after 24 weeks of gestation	2 (4)
Individual viable foetus at 24 weeks of gestation	1 (2)
Live birth event/delivery	19 (36)
Live birth event and individual baby (both given in article)	1 (2)
Individual living baby	1 (2)
Pregnancy > 28 weeks of gestation	1 (2)
Undefined	27 (51)

Table 4: Frequency (%) of definitions of 'live birth' in IVF trials reporting on this outcome in 2013-2014.

Ongoing pregnancy

S Table 14 shows the definitions of ongoing pregnancy, with around a third (13 studies, 33%) not providing any. Definitions were somewhat variable, with considerable differences in the gestational age required to declare that the pregnancy was ongoing.

Biochemical pregnancy

S Table 15 shows the definitions of biochemical pregnancy. These were almost universally defined on the basis of positive B-hCG tests, with variations arising from different cut-off values of the assay and different timings of testing.

4.5 Discussion

Our review confirms large-scale diversity in outcome reporting in IVF trials and suggests several areas of systematic methodological weakness in the evidence base. Over 800 combinations of numerator and denominator were reported, the majority of which were not used in more than one article. No single outcome measure appeared in a majority of studies. Subtly different definitions of numerators were employed, increasing the variety of reporting options even further. This affirms the concerns highlighted by the Core Outcomes in Women's Health (CROWN) Initiative who noted that a lack of common reporting standards was a hindrance to the synthesis of evidence (Khan, 2014). The recommendation set out in IMPRINT, that all infertility trials should report live birth and cumulative live birth, may go some way to address this matter. This review indicates that at present a minority of studies report live birth and few report cumulative live birth, although it was not common for studies to include multiple treatment cycles. The rates of reporting of live birth and other clinical outcomes are lower than was observed in a previous review of infertility trials, because the authors of that study required the reporting of a clinical outcome for inclusion (Dapuzzo, et al., 2011). Moreover, we have shown that where live birth is reported, a variety of denominators are used. Consequently, we suggest that the matter of combining outcomes with different denominators in meta-analysis warrants attention. We note that the proposition to have live birth as the primary outcome of all infertility trials would require all infertility trials to be powered to this end. This would rule out the possibility of smaller, explanatory trials, which may prove useful to the development of interventions. We suggest that procedural outcomes of treatment may be more appropriate for the evaluation of such trials. Live birth could still be reported, if not interpreted, and any intervention should ultimately be tested in confirmatory studies with live birth as the primary outcome. It is worth noting that using live birth as the primary outcome incurs practical disadvantages such as the need for a longer duration of follow up, which delays the release of clinical information and may be problematic in the eyes of funding bodies (Braakhekke, et al., 2014a). A compromise might be for journals to allow trial reports to be submitted for peer review with ongoing pregnancy results and, following acceptance of the manuscript, to require study authors to supply live birth results prior to publication. A consensus regarding what should constitute an ongoing pregnancy does not appear to exist at present however. We found a variety of definitions in use, with several studies describing pregnancies as ongoing prior to 12 weeks post transfer, contrary to the definition appearing in IMPRINT (Legro, et al., 2014). It was not usual for studies to contain an explicit description of live birth at all, and it was rarer still for studies to include a lower limit of gestation as part of the definition (such as the 20 weeks recommended by IMPRINT) (Legro, et al., 2014).Taking live birth as an example, we investigated denominator selection in more detail and found evidence that RCT methodology remains widely misunderstood by researchers and peer reviewers. Of those reporting live birth rates, a third of studies used the subgroup of patients achieving oocyte retrieval or embryo transfer as the denominator, rather than the set of all patients who were randomised earlier in the treatment process. The implications of this analytic strategy are more severe than just a loss of power. Randomised trials represent the gold standard in treatment evaluation due to the fact that random allocation to interventions ensures a balance over confounding factors. When outcomes are reported in subgroups of patients who reached a certain stage of the treatment process, and this does not coincide with the original randomised cohort, the balance is not preserved (Hirji and Fagerland, 2009, Yusuf, et al., 1991). Accordingly, any observed differences in outcome may be due to differences in prognostic characteristics rather than treatment effects. The comparative groups are particularly likely to differ when patients with certain characteristics are more or less likely to have a successful stimulation response or to achieve transfer in one arm of the trial (Hirji and Fagerland, 2009). Belief in the existence of such differential effects of treatment is the cornerstone of personalised IVF (Dewailly, et al., 2014, La Marca and Sunkara, 2014, Nelson, 2013). We expect that the issue will be more severe the greater the number of participants excluded, although this requires investigation in future simulation studies. The percentage of participants excluded in this sample tended to be less than 10%. A simple strategy to avoid this issue is that used by the Cochrane Gynaecology and Fertility Group, which is to define those participants for whom treatment has failed prior to embryo transfer as having an unsuccessful response. We also note that while it is valid to analyse results per transfer or per oocyte retrieval whenever patients have been randomised at this stage of treatment, reporting outcomes per cycle started may be more relevant to patients deciding whether or not to undertake IVF (Heijnen, et al., 2004). It may be argued that pragmatic effectiveness studies should therefore randomise prior to the start of the cycle (Mastenbroek and Repping, 2014). Other examples of statistical naiveté were identified. Some studies reported live birth per embryo transferred, which is problematic since embryos are not statistically independent and the outcome is defined at the level of the patient, rather than of the embryo (Vail and Gardener, 2003). Other studies randomly divided each patient's oocytes or embryos between intervention arms and compared the clinical outcomes between groups of patients who happened to have embryos transferred from only one of the arms. This is not a valid comparison, and may reflect the influence of initiatives promoting the reporting of clinical endpoints in all studies. We also suggest that the tendency to report myriad outcomes carries implications of false effect discovery due to multiple testing and selective emphasis or reporting. In theory, the specification of a primary outcome should offer some protection against these concerns, although in the absence of prospective trial registration there is no guarantee that the primary outcome has been selected in advance (Chan, et al., 2004). Moreover, these matters are particularly problematic given the fact that any outcome can be constructed in a variety of ways using the building blocks available combined with the strong emphasis on statistical significance in these trials. Outcome reporting bias would appear to represent an ungovernable potential source of bias in this field given that such a plethora of outcome measures are acceptable to peer reviewers.

Our study has limitations. This review was not comprehensive, as we restricted our sample to English-language publications in peer-reviewed journals. It is not clear however that publication bias represents a concern for a review of outcomes, as the accessibility of any particular study may not be related to the outcome measures used. The subgroup analysis of trials reporting live birth was not prespecified. It should also be noted that the categorisation scheme presented here is entirely arbitrary and was not prospectively designed; another review team likely would have made different decisions relating to the simplification and presentation of the outcome measures. The exact frequencies we present should be interpreted with this in mind. We believe that our conclusions are not contingent upon our particular scheme. Finally, due to the practice of reporting a trial's results across multiple publications, a small number of included articles may have been secondary publications reporting on particular secondary outcomes. Strictly speaking, the article, rather than the trial, is the unit of analysis in this review. We would suggest that it is appropriate to include these publications, as the decision to exclude them would omit reported outcomes where investigators had split results across two publications.

This is the first review to fully detail the outcomes reported across IVF trials. A previous review restricted their search to highly ranked journals and to studies reporting clinical outcomes (Dapuzzo, et al., 2011). This was suitable for the authors' aims of highlighting inconsistency in defining outcomes and underreporting of adverse events. It does not permit the prevalence of each outcome to be calculated however. Additionally, we note that high quality of reporting in all journals, not just the best, is a prerequisite for systematic review, where there is a need to identify all trials (although this would also include older studies, which we have not considered here). A second review found modest rates of reporting of neonatal and maternal outcomes in reproductive medicine trials (Braakhekke, et al., 2014b). However, that study restricted focus to outcomes in these two categories and only included trials appearing in Cochrane reviews. Accordingly, the results do not give a complete or representative picture of the current state of outcome reporting in IVF trials. There is massive diversity in numerator and denominator selection in IVF trials due to its multistage nature, and this causes methodological frailty in the evidence base. Existing efforts to improve the situation are certainly useful, although we would urge that future extensions to these projects include guidance on the definition and use of denominators as well as numerators and acknowledge that clinical outcomes may not be appropriate for early phase studies.

ACKNOWLEDGEMENTS

We would like to thank James Duffy, Lesley-Anne Carter, Antonia Marsden and two anonymous peer reviewers for helpful comments regarding an earlier draft of the manuscript.

CONFLICT OF INTEREST

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: JW is funded by a Doctoral Research Fellowship from the National Institute for Health Research (DRF-2014-07-050), supervised by AV and SR; The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. JW declares that publishing research benefits his career. DRB is funded by the NHS as Scientific Director of a clinical IVF service. AV and JW are statistical editors of the Cochrane Gynaecology and Fertility Group, although the views expressed here do not necessarily represent those of the group; no other relationships or activities that could appear to have influenced the submitted work.

4.6 References for Chapter 4.

Abdalla HI, Bhattacharya S, Khalaf Y. Is meaningful reporting of national IVF outcome data possible? *Hum Reprod* 2010;25: 9-13.

Blazeby J, Altman DG, Clarke M, Gargon EA, Williamson PR. Core outcome sets and the COMET (core outcome measures in effectiveness trials) initiative; improving the efficiency and value of the research process. *Qual Life Res* 2012;21: 19-20.

Braakhekke M, Kamphuis EI, Dancet EA, Mol F, van der Veen F, Mol BW. Ongoing pregnancy qualifies best as the primary outcome measure of choice in trials in reproductive medicine: an opinion paper. *Fertil Steril* 2014a;101: 1203-1204.

Braakhekke M, Kamphuis EI, Van Rumste MM, Mol F, Van Der Veen F, Mol BW. How are neonatal and maternal outcomes reported in randomised controlled trials (RCTs) in reproductive medicine?. *Hum Reprod* 2014b;29(6):1211-7.

Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;171: 735-740.

Clarke M, Williamson PR. Core outcome sets and systematic reviews. *Syst Rev* 2016;5: 11.

Dapuzzo L, Seitz FE, Dodson WC, Stetter C, Kunselman AR, Legro RS. Incomplete and inconsistent reporting of maternal and fetal outcomes in infertility treatment trials. *Fertil Steril* 2011;95: 2527-2530.

Dewailly D, Andersen CY, Balen A, Broekmans F, Dilaver N, Fanchin R, Griesinger G, Kelsey TW, La Marca A, Lambalk C *et al.* The physiology and clinical utility of anti-Mullerian hormone in women (vol 20, pg 370, 2014). *Hum Reprod Update* 2014;20: 804-804.

Garrido N, Bellver J, Remohi J, Simon C, Pellicer A. Cumulative live-birth rates per total number of embryos needed to reach newborn in consecutive in vitro fertilization (IVF) cycles: a new approach to measuring the likelihood of IVF success. *Fertil Steril* 2011;96: 40-46.

Germond M, Urner F, Chanson A, Primi M-P, Wirthner D, Senn A. What is the most relevant standard of success in assisted reproduction? The cumulated singleton/twin delivery rates per oocyte pick-up: the CUSIDERA and CUTWIDERA. *Hum Reprod* 2004;19: 2442-2444.

Griesinger G. Beware of the 'implantation rate'! Why the outcome parameter 'implantation rate' should be abandoned from infertility research. *Hum Reprod* 2016;31: 249-251.

Heijnen E, Macklon NS, Fauser B. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod* 2004;19: 1936-1938.

Hirji KF, Fagerland MW. Outcome based subgroup analysis: a neglected concern. *Trials* 2009;10.

Khan K. The CoRe Outcomes in WomeN's Health (CROWN) Initiative: Journal Editors Invite Researchers to Develop Core Outcomes in Women's Health. *Neurourol Urodyn* 2014;33: 1176-1177.

Khan K,. The CROWN Initiative: journal editors invite researchers to develop core outcomes in women's health. *BJOG* 2014;121: 1181-1182.

La Marca A, Sunkara SK. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Hum Reprod Update* 2014;20: 124-140.

Legro RS, Wu X, Barnhart KT, Farquhar C, Fauser BC, Mol B. Improving the reporting of clinical trials of infertility treatments (IMPRINT): modifying the CONSORT statementdaggerdouble dagger. *Hum Reprod* 2014;29: 2075-2082.

Legro RS, Wu X, Scientific C, Barnhart KT, Farquhar C, Fauser BC, Mol B. Improving the reporting of clinical trials of infertility treatments (IMPRINT): modifying the CONSORT statement. *Hum Reprod* 2014;29: 2075-2082.

Mastenbroek S, Bossuyt PMM, Heineman MJ, Repping S, van der Veen F. Comment 1 on Staessen et al. (2004). Design and analysis of a randomized controlled trial studying preimplantation genetic screening. *Hum Reprod* 2005;20: 2362-2363.

Mastenbroek S, Repping S. Preimplantation genetic screening: back to the future. *Hum Reprod* 2014;29: 1846-1850.

Min JK, Breheny SA, MacLachlan V, Healy DL. What is the most relevant standard of success in assisted reproduction? The singleton, term gestation, live birth rate per cycle initiated: the BESST endpoint for assisted reproduction. *Hum Reprod* 2004;19: 3-7.

Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ* 2001;165: 1339-1341.

Nelson SM. Biomarkers of ovarian response: current and future applications. *Fertil Steril* 2013;99: 963-969.

Vail A, Gardener E. Common statistical errors in the design and analysis of subfertility trials. *Hum Reprod* 2003;18: 1000-1004.

Van Voorhis BJ. Clinical practice. In vitro fertilization. *The N Engl J Med* 2007;356: 379-386.

Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical-Trials. *JAMA* 1991;266: 93-98.

4.7 Supplementary material for Chapter 4.

Numerator	**Denominator**	**No (%) of studies reporting item.**
Live birth or take home baby		52 (37% of studies)
	per cycle started (or earlier)	8 (6%)
	per patient achieving trigger	3 (2%)
	per oocyte retrieval	6 (4%)
	per embryo transferred	8 (6%)
	per transfer	21 (15%)
	par patient w/ sufficient embryos	5 (4%)
	unclear denominator	5 (4%)
Cumulative live birth		7 (5% of studies)
	per course of treatment started	4 (3%)
	time to pregnancy leading to live birth	2 (1%)
Preterm birth		4 (3% of studies)
	per baby	3 (2%)

S Table 1: Live birth outcomes reported by more than one study. Frequency (%) of studies reporting each outcome.

Numerator	Denominator	No (%) of studies reporting item.
Cycle cancellation		18 (13% of studies)
	per cycle started (or earlier)	13 (9%)
Total gonadotropin dose		39 (27% of studies)
	per cycle started (or earlier)	21 (15%)
	per oocyte retrieval	3 (2%)
	per transfer	5 (4%)
	unclear	4 (3%)
Good quality oocyte		3 (17%)
	per oocyte	2 (1%)
Mature oocyte		8 (6%)
	per oocyte	8 (6%)
Number of oocytes		65 (46%)
	per cycle started (or earlier)	28 (20%)
	per oocyte retrieval	9 (6%)
	per patient achieving fertilisation	9 (6%)
	per patient achieving trigger	9 (6%)
	per patient with sufficient follicles	9 (6%)
	per transfer	5 (4%)
	unclear	5 (4%)
Number of mature oocytes		37 (23%)
	per batch of oocytes	2 (1%)
	per cycle started (or earlier)	17 (12%)
	per oocyte retrieval	4 (3%)
	per patient achieving fertilisation	2 (1%)
	per patient achieving trigger	2 (1%)
	per transfer	3 (2%)
	unclear	2 (1%)
Number of mature oocytes/number of oocytes		8 (2%)
	per cycle started (or earlier)	3 (2%)
Number of degenerative oocytes		2 (1%)
	per cycle started (or earlier)	2 (1%)
Number of immature oocytes		2 (1%)
	per cycle started (or earlier)	2 (1%)
Number of follicles of sufficient size		14 (8%)
	per cycle started (or earlier)	7 (5%)
	per oocyte retrieval	2 (1%)
	unclear	2 (1%)

S Table 2: Stimulation outcomes reported by more than one study. Frequency (%) of studies reporting each outcome. Continued in S Table 3

Numerator	**Denominator**	**No (%) of studies reporting item.**
Number of inseminated oocytes		3 (2% of studies)
	per oocyte retrieval	2 (1%)
Oocyte retrieval achieved		8 (6%)
	per cycle started (or earlier)	5 (4%)
Oocyte retrieved		2 (1%)
	per follicle	2 (1%)
Stimulation duration		37 (26%)
	per cycle started (or earlier)	19 (13%)
	per oocyte retrieval	
		3 (2%)
	per patient achieving downregulation	2 (1%)
	per transfer	6 (4%)
	unclear	3 (2%)
Survival		4 (3%)
	per warmed oocyte	3 (2%)

S Table 3: Continuation of S Table 2. Stimulation outcomes reported by more than one study. Frequency (%) reporting each outcome

Numerator	**Denominator**	**No (%) of studies reporting item.**	**Numerator**	**Denominator**	**No (%) of studies**
Blastocyst		7 (5%)		unclear	22(16%)
	per 2PN embryo	2 (1%)	*Frozen embryo*		4 (3%)
	unclear	2 (1%)		per embryo obtained	2 (1%)
Blastocyst on day 5		2 (1%)	*Good quality embryo obtained*		22 (15%)
	per embryo obtained	2 (1%)		per cleavage stage embryo	2 (1%)
Cleavage		16 (11%)		per embryo obtained	6 (4%)
	per embryo obtained	7 (5%)		per oocyte	2 (1%)
	unclear	3 (2%)		unclear	8(6%)
Number of cleaved embryos		7 (5%)	*Good quality embryo on day 5*		3 (2%)
	per cycle started (or earlier)	2 (1%)		per embryo	3 (2%)
	unclear	2 (1%)	*Number of good quality embryos obtained*		17 (12%)
Embryo quality		11 (8%)		per cycle started (or earlier)	5 (4%)
	per embryo obtained	2 (1%)		per oocyte retrieval	2 (1%)
	per embryo transferred	2 (1%)		per transfer	3(2%)
	unclear	2 (1%)		unclear	3(2%)
Fertilisation failure		5 (4%)	*Number of 2PN embryos*		9 (6%)
	per cycle started (or earlier)	3 (2%)		per cycle started (or earlier)	6 (4%)
Fertilisation		53 (37%)	*Number of cells*		3 (2%)
	per cycle started (or earlier)	5 (4%)		per embryo obtained	2 (1%)
	per inseminated oocyte	3 (2%)	*Number of transfers*		2 (1%)
	per mature oocyte	6 (4%)		per group (totals)	2 (1%)
	per oocyte	8 (6%)			
	per oocyte retrieval	3 (2%)			
	per transfer	3 (2%)			

S Table 4: Fertilisation outcomes reported by more than one study. Frequency (%) of studies reporting each outcome. Continued in S Table 5.

Numerator	Denominator	No (%) of studies reporting item.
Number of embryos obtained		27 (19%)
	per cycle started (or earlier)	8 (6%)
	per oocyte retrieval	3 (2%)
	per transfer	3 (2%)
	per transfer of embryos from one intervention arm	2 (1%)
	unclear	5 (4%)
Number of frozen embryos		20 (14%)
	per cycle started (or earlier)	3 (2%)
	per patient with sufficient embryos	3 (2%)
	per transfer	2 (1%)
	unclear	7 (5%)
Survival		9 (6%)
	per warmed embryo	4 (3%)

S Table 5: Continuation of S Table 4. Fertilisation outcomes reported by more than one study. Frequency (%) of studies reporting each outcome.

Numerator	**Denominator**	**No (%) of studies reporting item.**
Transfer achieved		12 (8%)
	per cycle started (or earlier)	6 (4%)
	per oocyte retrieval	3 (2%)
Transfer with good quality embryos achieved		2 (1%)
	per cycle started (or earlier)	2 (1%)
Day 3 transfer achieved		3 (2%)
	per transfer	3 (2%)
Day 5 transfer achieved		4 (3%)
	per transfer	2 (1%)
Difficulty of transfer		3 (2%)
	per transfer	2 (1%)
Single embryo transfer		4 (3%)
	per transfer	3 (2%)
Double embryo transfer		4 (3%)
	per transfer	3 (2%)

S Table 6: Transfer outcomes reported by more than one study. Frequency (%) of studies reporting each outcome. Continued in S Table 7

Numerator	Denominator	No (%) of studies reporting item.
Implantation		74 (52%)
	per embryo obtained	3 (2%)
	per embryo transferred	42 (30%)
	per embryo transferred in a transfer of embryos from one intervention arm	2 (1%)
	per patient achieving trigger	2 (1%)
	per transfer	7 (5%)
	unclear	19 (38%)
Number of embryos transferred		74 (52%)
	per cycle started (or earlier)	12 (8%)
	per group (totals)	7 (5%)
	per oocyte retrieval	7 (5%)
	per patient achieving trigger	2 (1%)
	per patient with sufficient embryos	7 (5%)
	per patient with sufficient follicles	2 (1%)
	per transfer	24 (17%)
	unclear	14 (10%)
Number of gestational sacs		2 (1%)
	per group (totals)	2 (1%)
Number of transfers		2 (1%)
	per group (totals)	2 (1%)
Insemination method		6 (4%)
	per cycle started (or earlier)	2 (1%)

S Table 7: Continuation of S Table 6. Transfer outcomes reported by more than one study. Frequency (%) of studies reporting each outcome.

Numerator	**Denominator**	**No (%) of studies reporting item.**
Multiple pregnancy		24 (17%)
	per biochemical pregnancy	2 (1%)
	per clinical pregnancy	11 (8%)
	per ongoing pregnancy	2 (1%)
	per oocyte retrieval	3 (2%)
	per transfer	2 (1%)
	per patient w/ sufficient embryos	2 (1%)
Multiple birth		6 (4%)
	per live birth event	4 (3%)
	per transfer	2 (1%)

S Table 8: Multiple birth and pregnancy outcomes reported by more than one study. Frequency (%) of studies reporting each outcome.

Numerator	Denominator	No (%) of studies reporting item.
Ectopic pregnancy		18 (13%)
	per chemical pregnancy	3 (2%)
	per clinical pregnancy	2 (1%)
	per cycle started (or earlier)	2(1%)
	per patient with sufficient embryos	3(2%)
	per transfer	4(3%)
Mild OHSS		4 (3%)
	per cycle started (or earlier)	2 (1%)
Moderate OHSS		5 (4%)
	per patient achieving trigger	2 (1%)
	per transfer	2 (1%)
Severe OHSS		6 (4%)
	per patient achieving trigger	2 (1%)
OHSS (unspecified severity)		25 (18%)
	per cycle started (or earlier)	11 (8%)
	per oocyte retrieval	5 (4%)
	per patient with sufficient follicles	2(1%)
	per transfer	2(1%)
Adverse events (general)		10 (7%)
	per cycle started (or earlier)	3 (2%)

S Table 9: Adverse events reported by more than one study. Frequency (%) of studies reporting each outcome.

Numerator	Denominator	No (%) of studies reporting item.
Admission to intensive care		2 (1%)
	per baby	2 (1%)
Birthweight		6 (4%)
	per baby	4 (3%)
Congenital abnormalities		6 (4%)
	per baby	4 (3%)
Gestational age		3 (2%)
	per baby	2 (1%)
Low birthweight		2 (1%)
	per baby	2 (1%)
Sex		2 (1%)
	per baby	2 (1%)

S Table 10: Postnatal outcomes reported by more than one study. Frequency (%) of studies reporting each outcome.

Numerator	Denominator	No (%) of studies reporting item.
AMH		5 (4%)
	per cycle started (or earlier)	2 (1%)
	per oocyte retrieval	2 (1%)
AFC		5 (4%)
	per cycle started (or earlier)	2 (1%)
DHEA		4 (3%)
	per cycle started (or earlier)	2 (1%)
FSH		13 (9%)
	per cycle started (or earlier)	3 (2%)
	per oocyte retrieval	4 (3%)
	per transfer	3 (2%)
LH		*16 (11%)*
	per cycle started (or earlier)	6 (4%)
	per oocyte retrieval	3 (2%)
	per transfer	2 (1%)
	unclear	3 (2%)
Estradiol		45 (32%)
	per cycle started (or earlier)	16 (11%)
	per oocyte retrieval	7 (5%)
	per patient achieving downregulation	2 (1%)
	per patient achieving fertilisation	2 (1%)
	per transfer	6 (4%)
	unclear	6 (4%)
Estradiol to progesterone ratio		3 (2%)
	per cycle started (or earlier)	2 (1%)
Gene expression		3 (2%)
	per cycle started (or earlier)	2 (1%)
Kidney-yin deficiency symptom score		2 (1%)
	per cycle started (or earlier)	2 (1%)
Progesterone		17 (12%)
	per cycle started (or earlier)	6 (4%)
	per oocyte retrieval	2 (1%)
	per transfer	2 (1%)
	unclear	2 (1%)
Testosterone		5 (4%)
	per cycle started (or earlier)	2 (1%)
Endometrial thickness		35 (25%)
	per cycle started (or earlier)	16 (11%)
	per oocyte retrieval	5 (4%)
	per patient with sufficient follicles	2 (1%)
	per transfer	6 (4%)
	unclear	2 (1%)

S Table 11: Other procedural outcomes reported by more than one study. Frequency (%) of studies reporting each outcome

Definition of clinical pregnancy	Number of studies (%)	Definition of clinical pregnancy	Number of studies (%)
Confirmed by ultrasound 6-7 weeks of gestation	1 (1)	>=1 gestational sac and heartbeat on ultrasound 7 weeks of gestation	1 (1)
Foetal echoes and pulsations on ultrasound	1 (1)	>=1 gestational sac and heartbeat on ultrasound 8 weeks of gestation	1 (1)
Foetal pole and heartbeat	1 (1)	>=1 gestational sac and heartbeat on ultrasound 8-12 weeks of gestation	1 (1)
Foetus with heartbeat 6 weeks of gestation	1 (1)	>=1 gestational sac at 6 weeks	1 (1)
>=1 gestational sac on ultrasound 14-21 days post positive B-hCG	1 (1)	>=1 gestational sac on ultrasound	2 (2)
>=1 gestational sac 2 weeks post positive B-hCG	1 (1)	>=1 gestational sac on ultrasound 21 days of gestation	1 (1)
>=1 gestational sac 5 weeks of gestation	1 (1)	>=1 gestational sac on ultrasound 3 weeks of gestation	1 (1)
>=1 gestational sac and heartbeat	1 (1)	>=1 gestational sac on ultrasound 4 weeks of gestation	3 (3)
>=1 gestational sac and heartbeat 4 weeks of gestation	2 (2)	>=1 gestational sac on ultrasound 4-5 weeks of gestation	1 (1)
>=1 gestational sac and heartbeat 4-6 weeks of gestation	1 (1)	>=1 gestational sac on ultrasound 6 weeks of gestation	1 (1)
>=1 gestational sac and heartbeat 7 weeks	1 (1)	>=1 gestational sac on ultrasound 7 weeks of gestation	1 (1)
>=1 gestational sac and heartbeat at 8-10 weeks of gestation	1 (1)	>=1 gestational sac on ultrasound 7-14 days post positive B-hCG test	1 (1)
>=1 gestational sac and heartbeat on ultrasound	4 (4)	>=1 gestational sac or embryonic pole 4 weeks post OPU	1 (1)
>=1 gestational sac and heartbeat on ultrasound at 7 weeks of gestation	1 (1)	>=1 gestational sac or heartbeat on ultrasound	1 (1)
>=1 gestational sac and heartbeat on ultrasound 3 weeks post positive B-hCG test	1 (1)	>=1 gestational sac or heartbeat on ultrasound 6 weeks of gestation	1 (1)
>=1 gestational sac and heartbeat on ultrasound 5 weeks of gestation	3 (3)	>=1 gestational sac with foetal echoes and heartbeat	1 (1)
>=1 gestational sac and heartbeat on ultrasound 5-6 weeks of gestation	1 (1)	>=1 gestational sac, embryo and heartbeat on ultrasound 2-4 weeks post positive B-hCG	1 (1)
>=1 gestational sac and heartbeat on ultrasound 6 weeks of gestation	2 (2)	Heartbeat 4-5 weeks post OPU	1 (1)

S Table 12: Frequency (%) of definitions of ‘clinical pregnancy’ in IVF trials reporting on this outcome in 2013-2014. Continued in S Table 13

Definition of clinical pregnancy	Number of studies (%)
Heartbeat on ultrasound	3 (3)
Heartbeat on ultrasound 2 weeks post positive B-hCG test	3 (3)
Heartbeat on ultrasound 3 weeks post positive B-hCG test	1 (1)
Heartbeat on ultrasound 30 days of gestation	1 (1)
Heartbeat on ultrasound 4 weeks of gestation	2 (2)
Heartbeat on ultrasound 4-5 weeks post OPU	1 (1)
Heartbeat on ultrasound 5 weeks of gestation	1 (1)
Heartbeat on ultrasound 5 weeks post positive B-hCG	1 (1)
Heartbeat on ultrasound 7 weeks of gestation	2 (2)
Heartbeat on ultrasound 7-8 weeks of gestation	1 (1)
Heartbeat on ultrasound 7-9 weeks of gestation	1 (1)
One or more chambers or definitive clinical gestational signs	1 (1)
Positive B-hCG and ultrasound confirmation	1 (1)
Positive B-hCG test 11 days of gestation and gestational sac on ultrasound	1 (1)
Positive B-hCG test 14 days of gestation and gestational sac on ultrasound	1 (1)
Positive B-hCG test 15 days of gestation	1 (1)
Positive B-hCG test 2 weeks of gestation	1 (1)
Positive B-hCG test and gestational sac and Heartbeat on ultrasound	1 (1)
Positive B-hCG test and gestational sac on ultrasound 2 weeks post positive	1 (1)
B-hCG test positive B-hCG test and ultrasound confirmation at 6 weeks of gestation	1 (1)
Ultrasound confirmation 3 weeks post positive B-hCG test	1 (1)
Ultrasound confirmation 4 weeks of gestation	1 (1)
Ultrasound confirmation 6 weeks of gestation	1 (1)
Ultrasound confirmation 6-7 weeks of gestation	1 (1)
Ultrasound confirmation 7 weeks or 12 weeks of gestation	1 (1)
Undefined	21 (21)

S Table 13: Continuation of S Table 12. Frequency (%) of definitions of ‘clinical pregnancy’ in IVF trials reporting on this outcome in 2013-2014.

Definition of ongoing pregnancy	Frequncy (%) of studies
>=1 foetus with heartbeat on ultrasound at 6 weeks	1 (3)
>=1 viable foetus 20 weeks of gestation	1 (3)
Developing embryo 12 weeks of gestation	1 (3)
Foetus with heartbeat 12 weeks of gestation	1 (3)
>=1 gestational sac 18 weeks post transfer	1 (3)
>=1 gestational sac and heartbeat 4 weeks post transfer	1 (3)
>=1 gestational sac and heartbeat on ultrasound 12 weeks	1 (3)
>=1 gestational sac and heartbeat on ultrasound 12 weeks of gestation	1 (3)
Heartbeat on ultrasound 20 weeks	1 (3)
Heartbeat on ultrasound 6 weeks of gestation	1 (3)
Heartbeat on ultrasound 7 weeks of gestation	1 (3)
Pregnancy 10 weeks post start of treatment	1 (3)
Pregnancy 12 weeks of gestation	5 (13)
Pregnancy 16 weeks of gestation	1 (3)
Pregnancy 20 weeks of gestation	2 (5)
Pregnancy 24 weeks of gestation	1 (3)
Pregnancy with heartbeat on ultrasound 8 weeks of gestation	1 (3)
Ultrasound confirmation 10-12 weeks of gestation	1 (3)
Ultrasound confirmation 22 weeks of gestation	1 (3)
Ultrasound confirmation 8-10 weeks	1 (3)
Uncomplicated pregnancy rate 12 weeks of gestation	1 (3)
Undefined	13 (33)

S Table 14: Frequency (%) of definitions of ‘ongoing pregnancy’ in IVF trials reporting on this outcome in 2013-2014

Definition of biochemical pregnancy	Number of studies (%)
B-hCG>10IU/L	1 (2)
B-hCG > 10IU/L 12 days post transfer	2 (4)
B-hCG > 10IU/L 2 weeks post transfer	1 (2)
B-hCG > 10IU/mL 14-16 days post insemination OR positive B-hCG test 21-23 days post OPU	1 (2)
B-hCG > 20IU/L	1 (2)
B-hCG > 20IU/L 14 days post OPU	1 (2)
B-hCG > 50IU/L 2 weeks post transfer	1 (2)
B-hCG >= 30IU/L 14 days post transfer	1 (2)
B-hCG >= 50IU/L 12 days post transfer	1 (2)
Positive B-hCG test	11 (22)
Positive B-hCG test 12 days post transfer	4 (8)
Positive B-hCG test 13-15 days post transfer	1 (2)
Positive B-hCG test 14 days post OPU	1 (2)
Positive B-hCG test 14 days post transfer	9 (18)
Positive B-hCG test 14, 16 and 21 days post OPU	1 (2)
Positive B-hCG test 15 days post OPU	2 (4)
Positive B-hCG test 15 days post transfer	1 (2)
Positive B-hCG test 16 days post transfer	1 (2)
Positive B-hCG test 18 days post OPU	1 (2)
Positive B-hCG test or urine pregnancy test	1 (2)
Positive B-hCG test without gestational sac on ultrasound	1 (2)
Positive hCG test	1 (2)
Pregnancy test	1 (2)
Undefined	4 (8)

S Table 15: Frequency (%) of definitions of ‘biochemical pregnancy’ in IVF trials reporting on this outcome in 2013-2014

Chapter 5. Developments in IVF warrant the adoption of new performance indicators for ART clinics, but do not justify the abandonment of patient-centred measures.

Journal article 3

Authors Jack Wilkinson, Stephen A Roberts, Andy Vail

Status Published in Human Reproduction

Reference Wilkinson, J., Roberts, S.A., Vail, A. (2017). "Developments in IVF warrant the adoption of new performance indicators for ART clinics, but do not justify the abandonment of patient-centred measures." Hum Reprod **32**(6): 1155-1159.

Contribution statement JW devised the idea for the article. JW extracted and analysed data. All authors devised the content of the manuscript, contributed to the interpretation of the data, and wrote and edited the manuscript.

Preamble This piece was a direct response to as announcement by the Human Fertilization and Embryology Authority (HFEA) indicating that they intended to change the headline performance indicator on their online Choose a Clinic facility to 'live birth event per embryo transferred'. The motivation for this choice is to disincentivise the transfer of multiple embryos in a single procedure, which may be harmful to the mother and offspring. We presented statistical arguments to show that this measure could be misleading to patients, and that it does not have a clear interpretation. In the article we suggest that we should not be looking to construct a single measure that evaluates (and incourages) both effectiveness and safety simultaneously. Instead, we suggested that a set of measures should be presented, and gave an example of one such set.

Outputs and Impact of the research Following the publication of this article, HFEA recently launched their new Choose a Clinic website. The arguments advanced in this piece have not been taken on board.

5.1 Abstract

Recent advances in embryo freezing technology together with growing concerns over multiple births have shifted the paradigm of appropriate IVF. This has led to the adoption of new performance indicators for ART clinics by national reporting schemes, such as those curated by the Society for Assisted Reproductive Technology (SART) and the Human Fertilisation and Embryology Authority (HFEA). Using these organisations as case studies, we review several outcome measures from a statistical perspective. We describe several denominators that are used to calculate live birth rates. These include cumulative birth rates calculated from all fresh and frozen transfer procedures arising from a particular egg collection or cycle initiation, and live birth rates calculated per embryo transferred. Using data from both schemes, we argue that all cycles should be included in the denominator, regardless of whether or not egg collection and fertilisation were successful. Excluding cancelled cycles reduces the impact of confounding due to patient characteristics but also removes policy and performance differences which we argue represent relevant sources of variation. It may be misleading to present prospective patients with essentially hypothetical measures of performance predicated on parity of ovarian stimulation and transfer policies. Although live birth per embryo has the advantage of encouraging single embryo transfer, we argue that it is prone to misinterpretation. This is because the likelihood of live birth is not proportional to the number of embryos transferred. We conclude that it is not possible to present a single measure that encompasses both effectiveness and safety. Instead, we propose that a set of clear, relevant outcome indicators is necessary to enable subfertile patients to make informed choices regarding whether and where to be treated.

KEYWORDS

IVF/ outcome measure / national data/ Live birth /embryo/cumulative live birth/ success rates/ multiple pregnancy/ safety/ OHSS

5.2 Introduction

In vitro fertilisation (IVF) is a financially and emotionally burdensome treatment which, for the majority of patients, will end in failure. Most subfertile patients seek information about their condition online (Haagen, et al., 2003, Rawal and Haddad, 2006, Talarczyk, et al., 2012). Meanwhile, direct to consumer advertising of assisted reproductive technologies (ART) is ubiquitous (Abusief, et al., 2007, Huang, et al., 2005, Wilkinson, et al., 2017). Clinics compete for patients, creating a strong incentive to selectively report success rates in a manner that presents their performance as superior. The situation is particularly troubling, as the multistage nature of IVF introduces an extensive menu of numerators (such as live birth, or various stages of pregnancy) and denominators (such as the started cycle, transfer procedure, or individual embryo transferred) for this purpose (Heijnen, et al., 2004, Wilkinson, et al., 2016, Wilkinson, et al., 2017). Since individual clinics have no incentive to collaborate to provide consistent reporting of success rates, it falls to national reporting schemes to meet this challenge.

Historically, national reporting schemes such as those curated by the Society for Assisted Reproductive Technology (SART) in the US or the Human Fertilisation and Embryology Authority (HFEA) in the UK have emphasised live birth outcomes calculated in all fresh treatment cycles started. However, the widespread adoption of frozen embryo transfer (FET) together with growing concerns over the rate of multiple births has led to registries supplementing or changing the measures that they use to evaluate IVF programmes. Although the challenge of providing relevant information while protecting patients remains the same wherever IVF is offered, different strategies have emerged in response. In the following article, we consider these strategies and their implications from a statistical perspective.

5.3 Case Study 1: Society for Assisted Reproductive Technology

SART now present the outcome ‘preliminary cumulative outcome per intended egg retrieval’ at the top of the performance report for each of their member clinics on their *Find a Clinic* facility (Society for Assisted Reproductive Technology, 2016). This includes live birth events arising from all fresh and frozen transfers of embryos resulting from a cycle. The emphasis on this cumulative numerator reduces the incentive to transfer multiple embryos in the initial fresh transfer, because it ensures that this practice is not rewarded over the safer, and potentially more successful (Roberts, et al., 2011), option of transferring one embryo at a time in a series of transfer procedures. Importantly, the denominator includes all *intended* egg retrievals, so that any cycles cancelled prior to egg collection are included as failed treatments. Beneath this, the report presents live birth event in the first transfer procedure for each intended egg collection (cycle started), live birth per frozen cycle started and live birth per patient, which includes the outcome of any treatments undertaken by a new patient starting treatment at the clinic in the reporting year. SART's approach then has been to introduce and emphasise outcome measures that take into account all of the stages of treatment undertaken by patients, from the start of ovarian stimulation to the outcome of any subsequent transfer procedures. Consequently, the chosen measures are both clear and relevant to potential patients.

5.4 Case Study 2: Human Fertilisation and Embryology Authority

Following an extensive consultation process (Human Fertilisation and Embryology Authority, 2014), HFEA have announced changes to the way they report success rates through their online *Choose a Fertility Clinic* facility, a beta version of which is currently publically available (Human Fertilisation and Embryology Authority, 2016). The headline figure now presented for each clinic is 'live birth event per embryo transferred'. This counts birth events arising from each transfer procedure in the numerator, but increases the denominator by one for each individual embryo transferred to a patient (Abdalla, et al., 2010). Consequently, there is a penalty for multiple embryo transfer. If twins result from a double embryo transfer, live birth event per embryo transferred is ½ = 0.5. Patients who do not undergo a transfer procedure are excluded from the calculation. Beneath this, HFEA present 'cumulative live birth event per egg collection'. As for the cumulative birth measure reported by SART, this counts birth events resulting from the transfer of any embryos created from the oocytes obtained in a single egg collection. The two measures differ however in the fact that HFEA's version excludes patients who have their cycles cancelled prior to egg collection. Both of the measures emphasised in HFEA's new reporting standard therefore exclude a proportion of patients undergoing unsuccessful treatment.

Prospective patients must navigate to separate '*Detailed Statistics'* pages to find live birth events per cycle started, and cumulative birth rates per cycle started are not presented.

5.5 Consequences of excluding cancelled cycles

The numbers of patients who start ovarian stimulation but do not achieve egg collection or embryo transfer are nontrivial relative to the likelihood that treatment will be successful. Taking the 2014 figures from SART's most recent National Summary Report as an example, 9247 of 102,982 (9%) cycles started did not reach the egg collection stage, and of the 93,730 collections that do take place, 7188 (8%) had no embryos available for transfer (Society for Assisted Reproductive Technology, 2016). Outcome measures that exclude failed cycles effectively assume successful ovarian stimulation and fertilization and therefore exaggerate the chance of success for prospective patients.

Proponents of live birth event per embryo transferred have argued that the exclusion of cancelled cycles is actually an advantage of using the measure (Abdalla, et al., 2010). In particular, it is argued that cycle cancellation is largely driven by the prognostic characteristics of patients. Removing the initial stages of treatment from consideration removes not only much of the confounding due to differences in patient characteristics between centres, but also differences due to the variety of clinic embryo transfer policies. The strength of live birth event per embryo transferred as a measurer, it is claimed, is that it compares clinics purely in terms of the quality of the embryos they produce in their labs (Abdalla, et al., 2010).

This line of argument can be challenged. First, there is the matter of what should and should not be controlled for when comparing treatment programmes. It is indeed desirable to take differences in patient characteristics into account, lest confounding by indication obfuscate genuine differences in performance (Walker, 1996). It is less desirable to control for relevant differences in the treatment programmes themselves however, since it is these differences that account for much of the variation in success rates between clinics. Secondly, differences in patient characteristics may influence the uterine environment in addition to the stimulation and fertilisation stages (Roberts, et al., 2010), so are not fully resolved by the embryo selection process.

The variation in cancellation rates is demonstrated in Figure 8, which displays the proportions of cycles cancelled prior to oocyte retrieval in 62 clinics in the UK. The search strategy is contained in the supplementary material for this article, and the data are available at goo.gl/lKxQwz . This isn't an exhaustive list of UK clinics, but is a sample of sufficient size to illustrate the point at hand. The left panel shows the data for patients of all ages, ordered by cancellation rate. There is clear variation in cancellation rates. The right hand panel shows the cancellation rates for patients under 35 years of age, using the same ordering as the left hand panel. By limiting our attention to patients under 35, we reduce (but do not remove) variation due to patient characteristics. Moreover, younger patients will have higher ovarian reserve on average, and it is in this group that there is greatest scope for variation in ovarian response according to the clinic stimulation strategy (Fleming, et al., 2013). Clearly, the right hand panel shows that there remains considerable variation in cancellation rates, even after allowing for the increase in uncertainty arising from reduced sample sizes. It is artificial to pretend that these differences do not exist by choosing outcome measures that eliminate them from consideration.

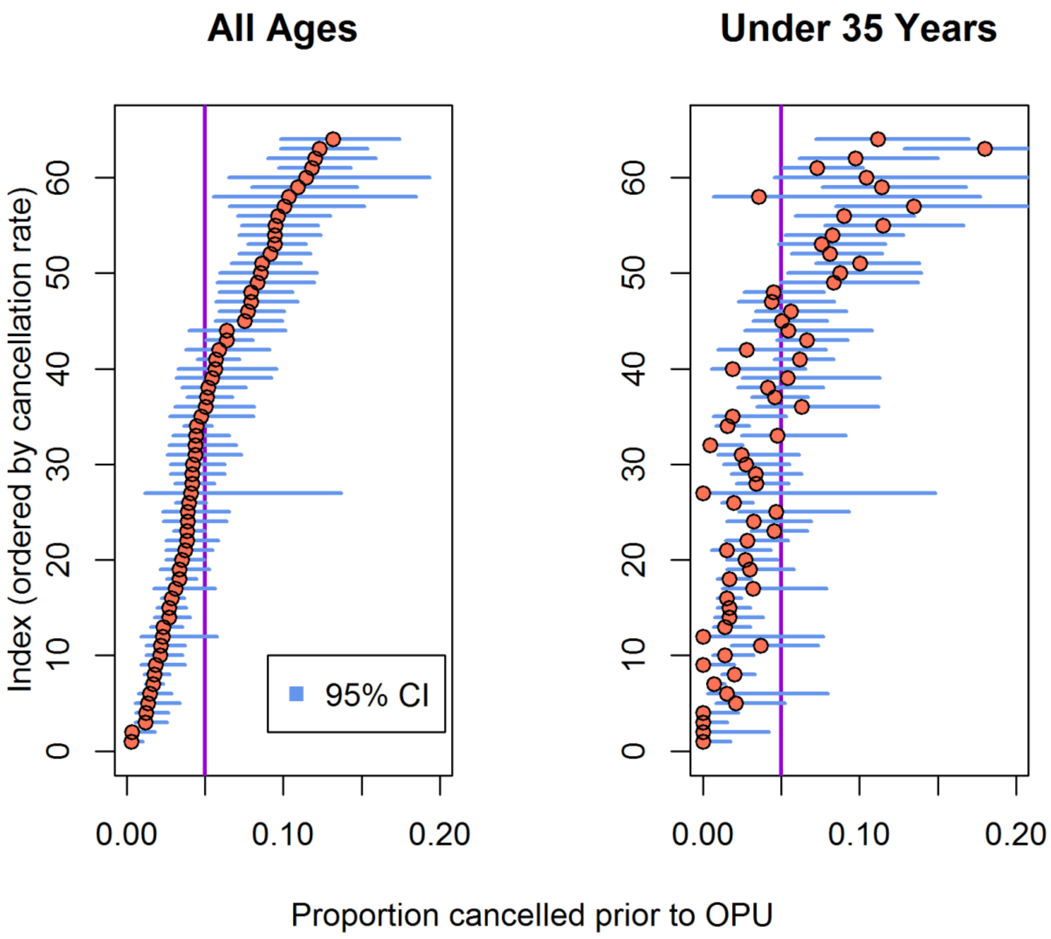


Figure 8: Proportion of cycles using own eggs cancelled prior to OPU in 64 clinics in the United Kingdom for the year ending Q2 2014. Vertical line is the pooled mean in the overall sample estimated using a random effects meta-analysis. Data extracted from HFEA

5.6 How to interpret live birth per embryo transferred?

Live birth per embryo transferred, aims to measure clinic performance in embryo transfer, were all other differences in treatment to be removed from consideration. This may have value for regulators and commissioners of services. However for patients this does not represent a comparison between actual treatment outcomes. Instead, it provides pseudo-information about a state of affairs that doesn't actually exist. Aside from this, there are other obstacles to interpretation, arising from the fact that the outcomes of multiple embryos transferred to the same patient are not statistically independent. For example, suppose we have a 'live birth event per embryo transferred' rate of 26% for some clinic. A patient might look at this and think that if they have two embryos transferred then the chance of success will be 52%, indicating that it would be more likely than not that they would have a baby. It is wrong, but not unreasonable, for a patient to arrive at this conclusion, because any statistic presented 'per unit' implies that if you have more units you will have proportionally more events. This is clearly not the case for live birth events per embryo transferred, because whether or not an embryo implants is partially determined by factors that have nothing to do with the embryos themselves. This immediately presents a serious concern, as it presents double embryo transfer as an attractive option. A technical consequence of this statistical dependency is that it is not possible to calculate valid confidence intervals on the basis of the total number of births and the total number of embryos transferred (Vail and Gardener, 2003).

5.7 Do the measures encourage safe treatments?

One motivation behind such measures as 'live birth per embryo transferred' and 'cumulative live birth per egg collection' is to promote patient safety by disincentivising multiple embryo transfer. However, while multiple births are a serious concern, they do not represent the only potential adverse consequence of treatment. Excessive response to ovarian stimulation is associated with increased risk of ovarian hyperstimulation syndrome (OHSS) (Steward, et al., 2014) and of preterm birth and low birthweight (Sunkara, et al., 2015). Meanwhile, analyses of large national databases have revealed that many stimulation cycles result in the retrieval of more than 15 oocytes; the figure has been estimated as 17% in the UK (Sunkara, et al., 2011) and 28% in the US (Steward, et al., 2014). Outcome measures that exclude the stimulation phase do not penalise, and may actually encourage, harsh stimulation strategies as clinics pursue larger oocyte yields to permit multiple frozen transfer procedures.

The lesson here is that we should not attempt to devise a single outcome measure that quantifies a clinic's safety and effectiveness because this approach removes the ability to consider each of these factors separately (Braakhekke, et al., 2015). Instead, it would be more appropriate to require adverse events to be explicitly reported, so that this information can be taken into account by potential patients. In Table I, we give an example of a set of outcomes covering effectiveness and safety which could be adopted by national reporting schemes.

5.8 Conclusion

Developments to IVF have created new challenges for national reporting schemes. A major motivation is to protect patients from unsafe treatments. One strategy to address these issues is to introduce innovative outcome measures designed to discourage superficially attractive, but ultimately detrimental practices. However, some of these measures encourage safe practices at the expense of providing clear, relevant information to couples. This is neither desirable nor necessary. A prospective patient must decide whether and where to undergo treatment prior to the start of ovarian stimulation. Pertinently, psychological and physical burden of treatment (Troude, et al., 2014, Verberg, et al., 2008) and of ovarian stimulation in particular (Verberg, et al., 2008) have been identified as predictors of treatment discontinuation. Outcome measures that ignore the stimulation phase therefore do not assess clinic performance in the dimensions that are important to many patients.

Attempts to adjust for differences in patient characteristics through choice of outcome measure are misguided. For example, the measure ‘live birth event per embryo transferred’ throws the baby out with the bathwater, by not only reducing the impact of patient characteristics but also removing relevant policy and performance variation. The ability to quantify relevant variation is a prerequisite of a useful performance indicator (Bird, et al., 2005). We would recommend either presenting success rates that are statistically adjusted for key confounders, or presenting ‘headline’ results stratified according to these relevant prognostic variables. This raises concerns about small sample sizes within strata, but we would agree with the suggestion that results should be presented over longer periods of time to reduce the impact of random noise (Chetkowski, 2014).

The availability of independently validated clinic-level success rates is a potentially powerful resource for patients, and one that is denied to many prospective patients around the world who must rely upon clinics’ own advertising, which may be prone to reporting biases (Wilkinson, et al., 2017). This is not statistical pedantry; the outcomes presented are used by vulnerable people facing a potentially life-changing decision. It is therefore essential that this information remains relevant and easy to understand, so as not to unintentionally mislead prospective patients and thereby deny them the opportunity to make a truly informed choice. Emphasising a small set of several outcome measures may be one way to achieve this.

Acknowledgements

We would like to thank two anonymous peer reviewers, whose comments substantially improved the article.

Conflict of interest statement

JW is funded by a Doctoral Research Fellowship from the National Institute for Health Research (DRF-2014-07-050) supervised by AV and SAR. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. AV and JW are statistical editors of the Cochrane Gynaecology and Fertility Group. SR is a statistical editor for Hum Reprod. AV has received grants from the Human Embryology and Fertilisation Authority, outside the submitted work. JW also declares that publishing peer-reviewed articles benefits his career.

5.9 References for Chapter 5

Abdalla HI, Bhattacharya S, Khalaf Y. Is meaningful reporting of national IVF outcome data possible? *Hum Reprod* 2010;25: 9-13.

Abusief ME, Hornstein MD, Jain T. Assessment of United States fertility clinic websites according to the American Society for Reproductive Medicine (ASRM)/Society for Assisted Reproductive Technology (SART) guidelines. *Fertil Steril* 2007;87: 88-92.

Bird SM, David C, Farewell VT, Harvey G, Tim H, Peter C. Performance indicators: good, bad, and ugly. J R Stat Soc Ser A Stat Soc 2005;168: 1-27.

Braakhekke M, Kamphuis EI, Mol F, Norman RJ, Bhattacharya S, van der Veen F, Mol BWJ. Effectiveness and safety as outcome measures in reproductive medicine. *Hum Reprod* 2015;30: 2249-2251.

Chetkowski RJ. Consumer-friendly reporting of in vitro fertilization outcomes. *Fertil Steril* 2014;101: e7.

Fleming R, Broekmans F, Calhaz-Jorge C, Dracea L, Alexander H, Andersen AN, Blockeel C, Jenkins J, Lunenfeld B, Platteau P *et al.* Can anti-Mullerian hormone concentrations be used to determine gonadotrophin dose and treatment protocol for ovarian stimulation? *Reprod Biomed Online* 2013;26: 431-439.

Haagen EC, Tuil W, Hendriks J, de Bruijn RPJ, Braat DDM, Kremer JAM. Current Internet use and preferences of IVF and ICSI patients. *Hum Reprod* 2003;18: 2073-2078.

Heijnen E, Macklon NS, Fauser B. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod* 2004;19: 1936-1938.

Huang JY, Discepola F, Al-Fozan H, Tulandi T. Quality of fertility clinic websites. *Fertil Steril* 2005;83: 538-544.

Human Fertilisation and Embryology Authority. Information for Quality Consultation. http://www.hfea.gov.uk/9633.html. Accessed 12/01/17.

Human Fertilisation and Embryology Authority. Choose a clinic (beta version). https://beta.hfea.gov.uk/. Accessed 12/01/17.

Rawal N, Haddad N. Use of Internet in infertility patients. *The Internet Journal of Gynecology and Obstetrics* 2006;5.

Roberts SA, Hirst WM, Brison DR, Vail A, TowardSET Collaboration. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. *Hum Reprod* 2010;25: 2792-2802.

Roberts SA, McGowan L, Hirst WM, Vail A, Rutherford A, Lieberman BA, Brison DR, Collaboration T. Reducing the incidence of twins from IVF treatments: predictive modelling from a retrospective cohort. *Hum Reprod* 2011;26: 569-575.

Society for Assisted Reproductive Technology. Find a Clinic. http://www.sart.org/find_a_clinic/. Accessed 12/01/17.

Society for Assisted Reproductive Technology. National Summary Report. https://www.sartcorsonline.com/rptCSR_PublicMultYear.aspx?ClinicPKID=0. Accessed 12/01/17.

Steward RG, Lan L, Shah AA, Yeh JS, Price TM, Goldfarb JM, Muasher SJ. Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: an analysis of 256,381 in vitro fertilization cycles. *Fertil Steril* 2014;101: 967-973.

Sunkara SK, La Marca A, Seed PT, Khalaf Y. Increased risk of preterm birth and low birthweight with very high number of oocytes following IVF: an analysis of 65 868 singleton live birth outcomes. *Hum Reprod* 2015;30: 1473-1480.

Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. *Hum Reprod* 2011;26: 1768-1774.

Talarczyk J, Hauke J, Poniewaz M, Serdynska-Szuster M, Pawelczyk L, Jedrzejczak P. Internet as a source of information about infertility among infertile patients. *Ginekol Pol* 2012;83: 250-254.

Troude P, Guibert J, Bouyer J, de La Rochebrochard E, Grp D. Medical factors associated with early IVF discontinuation. *Reprod Biomed Online* 2014;28: 321-329.

Vail A, Gardener E. Common statistical errors in the design and analysis of subfertility trials. *Hum Reprod* 2003;18: 1000-1004.

Verberg MF, Eijkemans MJ, Heijnen EM, Broekmans FJ, de Klerk C, Fauser BC, Macklon NS. Why do couples drop-out from IVF treatment? A prospective cohort study. *Hum Reprod* 2008;23: 2050-2055.

Walker AM. Confounding by indication. *Epidemiology* 1996;7: 335-336.

Wilkinson J, Roberts SA, Showell M, Brison DR, Vail A. No common denominator: a review of outcome measures in IVF RCTs. *Hum Reprod* 2016;31: 2714-2722.

Wilkinson J, Vail A, Roberts SA. Direct-to-consumer advertising of success rates for medically assisted reproduction: a review of national clinic websites. *BMJ Open* 2017;7.

5.10 Supplementary material for Chapter 5.

Search method used to extract data for Figure 1.

- Used HFEA Choose a Fertility Clinic website (http://guide.hfea.gov.uk/guide/).
- Searched for clinics offering IVF and ICSI.
- Performed this search separately by each 'region' listed on the site.
- Under birth data, clicked 'take a closer look'
- Looked at IVF/ICSI data.
- Under 'Details cycles and cancellations' looked at Year ending 2014 Q2 (the most recent data), Treatment cycles IVF and ICSI, Age group All ages, Embryo source fresh embryos, patient's eggs.
- Extracted cycles started, no. of cycles reaching OPU stage, no. of cycles reaching transfer stage. Total number of embryos transferred during all cycles.
- Using the same settings, 'Pregnancies and live births per treatment cycle'
- Then went through again and extracted same data for under 35s.

- The search yielded 74 clinics. 8 of these had not been open long enough to have results for this period. One clinic had unusual results (with no cycles reaching OPU). After making inquiries with HFEA, we established that due to renovation work at the clinic, some cycles were started here but completed at another clinic. It was not possible to disentangle these cycles from the rest using the summary data available, and we excluded both clinics.

III. Modelling multistage IVF outcomes

Chapter 6. Methods for the development of stage-specific IVF models

A prerequisite of a holistic IVF model is to establish good representations of each stage. We noted in the literature review that previous multistage IVF models had treated the response variables as binary indicators denoting success or failure, wasting information and providing an unrealistic representation of the underlying processes. In this chapter we start by considering the problem of how to model the stimulation stage, before turning our attention to embryo culture. The success of the embryo culture stage can be measured by embryo quality. Several dimensions of embryo quality can be measured (number of cells, evenness of cells, degree of fragmentation) and we consider these as a multivariate response. We extend this by representing the response variables using a mixture of outcome types (continuous and ordinal), and consider the problem of how to jointly model these mixed variables. These initial forays into joint modelling will serve as a basis for our multistage models, when, in the next chapter, we extend this work to the modelling of sequential treatment stages with mixed responses measured at different levels of a multilevel data structure.

6.1 How to model responses at the stimulation phase?

A journal article answering a clinical research question about the stimulation stage of treatment is included in the results section of the thesis (Journal Article 4, Chapter 9). Here, we detail the methodological considerations behind our modelling of the stimulation stage of treatment, which is only briefly covered in the article.

6.1.1. Motivation

The literature review highlighted the stimulation phase of the IVF treatment process as requiring particular attention, with existing approaches to modelling IVF as a multistage treatment either treating the outcome of this phase as binary or omitting it entirely. It was suggested in section 1.3.4 that particular aspects of the stimulation response, specifically the number and quality of oocytes obtained, may be important determinants of overall treatment outcome. A framework capable of incorporating these aspects of the stimulation outcome might therefore be desirable.

We conducted an analysis of stimulation outcomes, using this to develop our understanding of the methodological issues represented by this stage. This analysis was focussed on answering a clinical research question relating to the effect of gonadotropin dose on stimulation response. However, in the context of the thesis, we were equally concerned with the methodological question of how best to model data from this phase.

6.1.2. The clinical research question: what are the sources of variation in ovarian response, and what are the implications for ovarian stimulation?

When commencing IVF or ICSI, the patient's ovaries may be stimulated by the administration of gonadotropins, with the goal being to obtain a sufficient yield of oocytes to permit the creation of several good quality embryos without triggering an excessive response (La Marca and Sunkara, 2014). Excessive response represents a risk to the patient and offspring, and may result in termination of, or delay to, treatment. To this end, the patient's progress is monitored by ultrasound during the stimulation phase, and the daily gonadotropin dose may be adjusted accordingly. If it appears that the response is likely to be excessive or unsuccessful in the sense of yielding no oocytes (eggs), then the cycle may be cancelled, so that subsequent stages of treatment cannot be enacted. A current theme in IVF research is the question of whether or not a starting dose of gonadotropin can be selected on the basis of prognostic markers in such a way so as to optimise the response (personalised ovarian stimulation).

In this study, our interest was to investigate the sources of variation in ovarian response to stimulation, and to determine the implications for personalised treatment.

6.1.3. Introduction to the dataset

The dataset used here contains information on patients undergoing ovarian stimulation prior to IVF or ICSI (injection of the eggs with sperm) at St Mary's Hospital Department of Reproductive Medicine between October 1st 2008 and 8th August 2012. For the present analysis we excluded patients undergoing surgical interventions and patients with polycystic ovaries, as the relationships between prognostic variables and treatment outcomes are known to differ in these particular subgroups so that the patterns of confounding are distinct.

In addition to the exclusions noted above, we also excluded six observations for which the initial dose of gonadotropin was recorded as zero. It was not possible to identify whether these were recorded in error, and if not what was meant by this. Subsequently, the dataset contained 1851 treatment cycles on 1430 patients. 1070 (75%) patients had one cycle, 306 (21%) had two, 56 (4%) had three and 1 (0%) had four. We refrain from giving further detail about the characteristics of patients here, since this is covered in detail in Journal Article 4 (Chapter 9).

6.1.4. Representing the response to stimulation

The question of how to represent the outcome of ovarian stimulation was discussed in section 1.3.4, where it was noted that the stimulation response is often dichotomised or otherwise categorised. Such approaches waste information. Instead, we opted to model the number of oocytes obtained from the stimulation phase. We were also interested in the quality of eggs retrieved. This is more difficult to model, as a good measure of quality is not available. However, eggs of patients undergoing ICSI are stripped prior to the injection of sperm, and it is recorded whether or not each egg has matured. We therefore also modelled the number of mature eggs in the ICSI subset, as a way to investigate the impact of changing dose on egg quality. We used the same model for this as was used for the analysis of egg count, and this is not discussed separately here.

6.1.5. Covariate selection

The variables to be included in the model were selected on the basis of the background knowledge of our clinical collaborators and the objectives of the study. These variables were age, antimullerian hormone (AMH, a marker of ovarian reserve), antral follicle count (AFC, another marker of ovarian reserve), initial dose of gonadotropin, stimulation regime (antagonist or long downregulation), protocol (old, v1, v2 or v3, v4), type of gonadotropin (HMG or rFSH), USOR practitioner, attempt number, BMI and cause of infertility. Although there were considerable amounts of missing data for BMI, it was included as it is believed to be predictive of overall treatment success, and there is interest in identifying the role it plays in this phase. The representation of age, AMH and AFC in the model was determined on the basis of exploratory analysis consisting of graphing each variable against egg count and log(egg count), and by comparing models featuring competing representations using AIC (Akaike's Information Criterion, a measure of fit that penalizes complexity in the model, Akaike, 1972). The candidate representations were linear, quadratic and cubic representations on both the original and log transformed scales, splines with varying numbers of knots and categorical representations with varying numbers of categories. Where alternative representations performed similarly, the decision over which form to use was made on the basis of clinical interpretability. As a result of this process, age was represented as a quadratic in the final analysis, AMH was log-transformed and AFC was categorised into 3 levels on the basis of quantiles. Initial dose of gonadotropin was represented as a categorical variable. This decision was made on the basis of the distribution of the doses and the desire to obtain an easily interpretable answer to the research question. A patient is given one of 12 starting doses (so that the initial doses appearing in the data do not constitute a continuous scale) and several of these were used very sparingly. Accordingly, some patients were combined into dose bands. Interactions between both regime and other variables and dose and other variables were considered, using likelihood ratio testing and graphing of the predictors against egg count within regime and dose categories. The relationships between variables and egg count appeared to be similar for patients treated under both regimes on the basis of graphical representations of the data. However, dose effect was allowed to vary with regime in the final analysis, owing to the observed significance of this interaction using a likelihood ratio test and the inherent plausibility of this relationship. No other interactions were included in the final analysis.

6.1.6. Model selection: lognormal vs count models

The outcome variable ‘egg count' is discrete and skewed (Figure 9). As such, linear regression of the untransformed count variable is unlikely to be appropriate. Alternative approaches would be to model the outcome using an appropriate distribution for count data or to use linear regression after applying a log transformation. Figure 9 shows the log transformed egg counts after first adding 0.5 in order to accommodate the small number of zero values. The count model might be preferred on a priori grounds, as it does not require the addition of a small constant in order to incorporate zero outcomes and maintains the discrete nature of the data. O’Hara & Kotze, (2010) present a simulation study suggesting that generalised linear models are generally preferable to transformations for count data. In order to investigate this point further, two models were compared using AIC and posterior predictive checks; a linear regression model with log(eggs + 0.5) as the outcome variable and a Poisson model incorporating a gamma-distributed random effect at the observation level to account for overdispersion. During this model development stage, only complete cases were included and covariates were represented as described above; an exception was BMI, which was not included in posterior predictive checking due to the missingness in the variable.

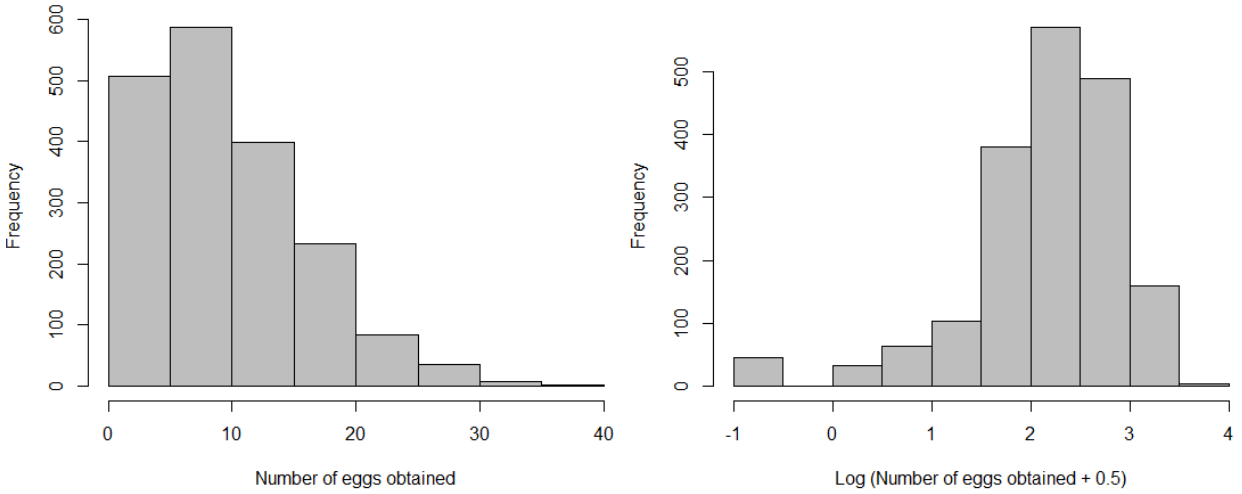


Figure 9: Histograms of number of eggs obtained (left) and log(number of eggs obtained+0.5) for 1851 IVF treatment cycles.

These developmental analyses did not account for the clustering of repeated cycles on the same patient. AIC was substantially lower for the log-normal model (2468 compared to 8591). However, given that the data have been altered by the addition of a constant in the case of the log-normal fit, it is unclear that comparison using AIC is strictly appropriate. Posterior predictive checks (Gelman, et al., 1996) might be more useful for the purpose of understanding the suitability of the models under consideration, and in particular the aspects in which they may be deficient. To perform posterior predictive model checking, both models were fit using markov chain monte carlo (MCMC) in order to obtain the joint posterior distribution of the model parameters. Weak priors were used so that the posterior was determined by the data. Specifically, in the count model, regression coefficients were assigned Normal (0, 1000^2) priors, and the random effects were assigned a Gamma(γ,γ) distribution with $\gamma \sim$ Gamma(0.001, 0.001). In the log-normal model, regression coefficients were again assigned Normal (0, 1000^2) priors. A uniform prior was considered appropriate for the standard deviation given that we do not have particular interest in this parameter here. For the log-normal model, three chains were run for 1000 iterations, the first half of which were discarded as burn-in. Convergence was assessed using traceplots and the Gelman-Rubin convergence statistic (Gelman and Rubin, 1992). The post burn-in draws from the three chains were pooled, resulting in 1500 draws from the posterior distribution. The procedure was similar for the count model; two chains were run for 10000 iterations each in order to ensure convergence of the random effects.

Once the posterior distributions had been obtained for the parameters of both models, predicted egg counts for patients in hypothetical replications of the study were simulated. These constitute the posterior predictive distributions of each model. By comparing the predictions generated by the models to the observed data, it is possible to consider the extent and manner in which the models are discrepant, indicating a lack of fit. Histograms of the observed egg counts and the egg counts from 19 datasets randomly selected from the posterior predictive distribution for the two models are shown in Figure 10 and Figure 11. Noting the differences in axes, the histograms from the overdispersed count model bear a closer resemblance to the observed egg counts. In particular, Figure 10 suggests that the lognormal model would occasionally predict egg counts that are much higher than are actually observed in the data, which does not appear to be the case for the overdispersed Poisson replications. Further insight can be gained into how well each model captures different aspects of the data by comparing summary statistics (often called test or discrepancy statistics) calculated from the observed data to the corresponding statistics calculated for each replicated dataset (Gelman, et al., 1996).

The maximum and minimum egg counts predicted by the models constitute relevant discrepancy statistics here. A caveat is that very high values do not appear in the observed data due to the fact that some cycles predicted to have excessive response are halted. Nonetheless, the biological plausibility of the values generated by the model can be considered. Figure 12 shows the maximum values from the predicted datasets compared to the maximum value of the observed egg counts, for both the lognormal and count models. It can be seen that the log-normal model systematically overpredicts the

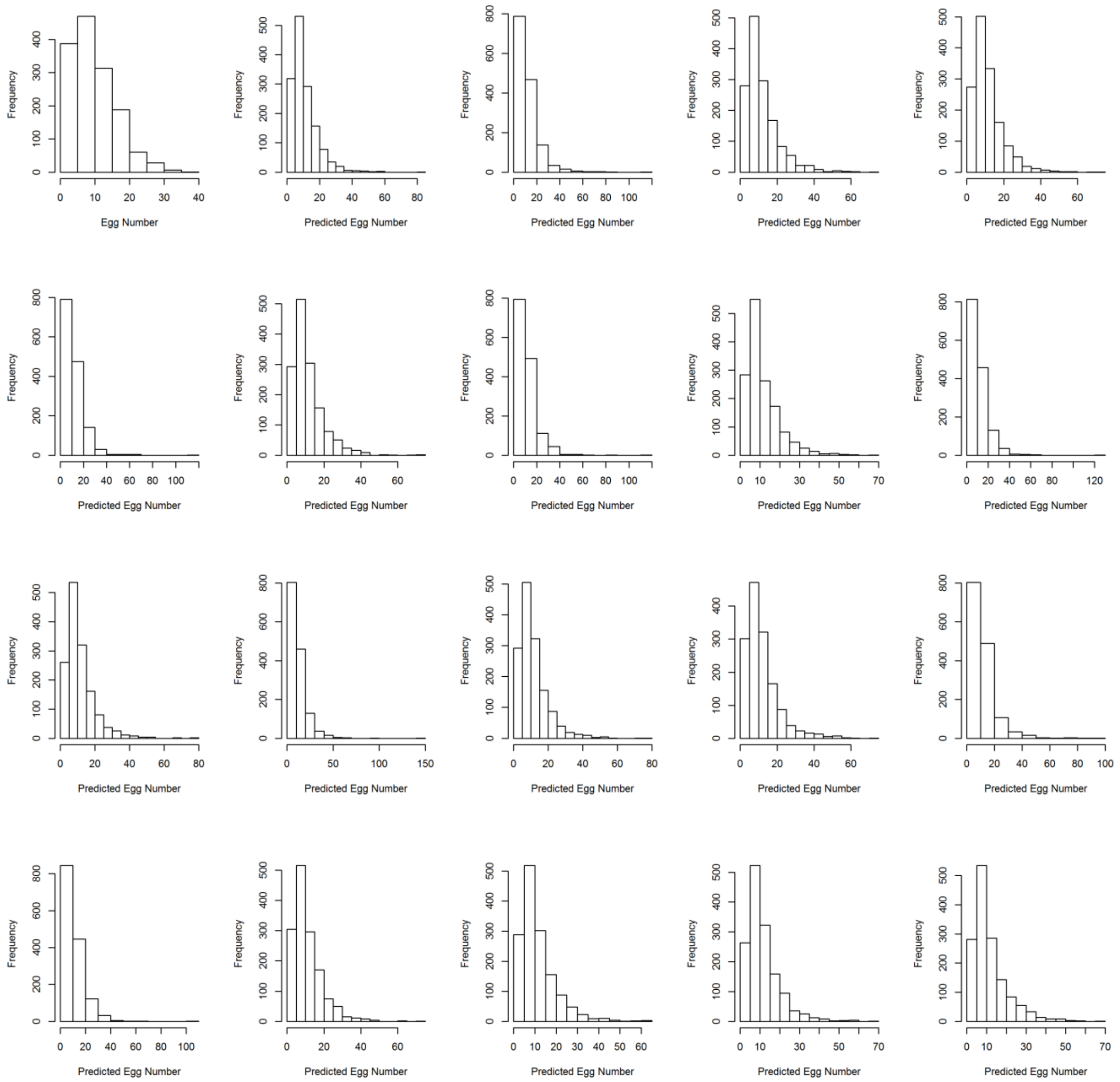


Figure 10: Observed egg counts (top left) and hypothetical replications drawn from the posterior predictive distribution of the lognormal model. Note differences in axes.

maximum egg count. The maximum in the observed data is 38 eggs. The range of maximum egg counts in the replicated log-normal datasets is 50 to 238, exceeding biological plausibility. Although the count model does tend to overpredict the maximum egg count, this is less extreme than is seen for the log-normal model. We would not rule out the count model on the basis of a Bayesian p-value, calculated as the proportion of replicated maximum egg counts exceeding the observed value (P = 0.86, where large or small values would suggest systematic discrepancy). Similarly, the log-normal model cannot produce counts of exactly zero, so the minimum egg count in the observed data is never predicted by the model (not shown). The median (IQR) minimum egg count in the replicated log-normal datasets is 0.85 (0.72 to 0.97). By contrast, all of the replicated datasets drawn from the count model have minimum values of zero.

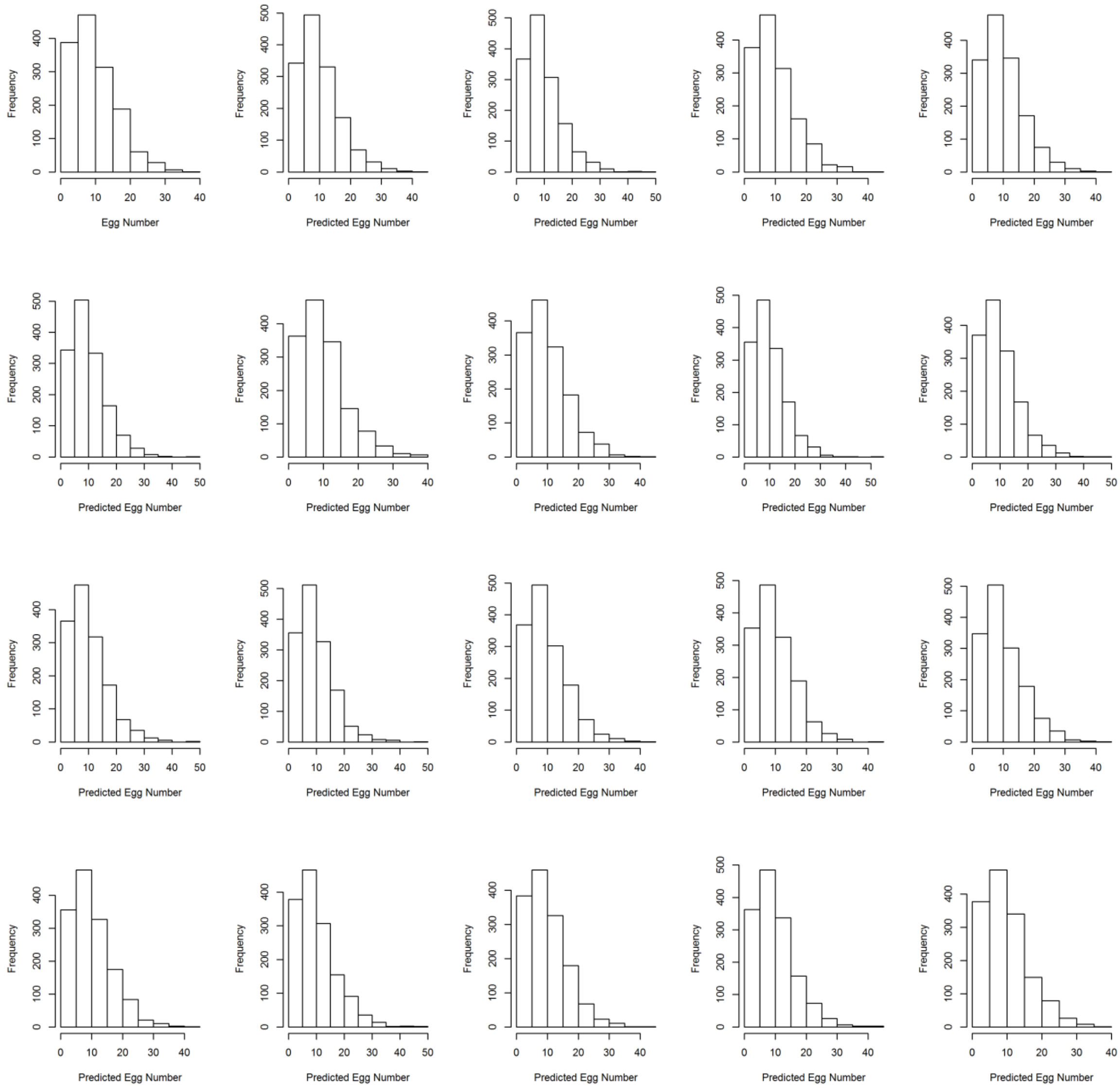


Figure 11: Observed egg counts (top left) and hypothetical replications drawn from the posterior predictive distribution of the overdispersed count model (note differences in axes).

Further posterior predictive checks of the mean, SD, median, IQR and the ratio of mean to variance were performed (not shown). These suggest that the count model is consistent with the observed data. The log-normal model captures the median and IQR well, but systematically overpredicts the mean and SD and underpredicts the mean/variance ratio. These checks suggest that the count model is preferable to the log-normal. Consequently, we gave precedence to count models in the final analysis, although we did also fit lognormal models in order to investigate the implications for inference.

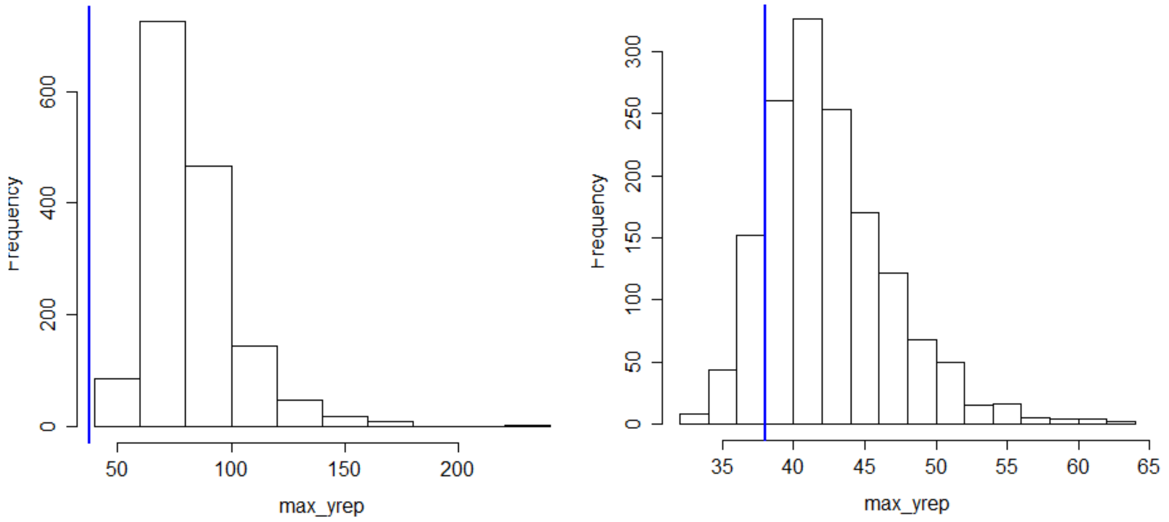


Figure 12: Posterior predictive checks of log-normal (left) and overdispersed count (right) models using the summary statistic max(egg count). The maximum counts from replicated datasets drawn from each model are displayed. The vertical blue line shows the maximum egg count in the data. Note that the plots have different scales.

6.1.7. Multiple imputation of missing values

Here, the multiple imputation of missing values is briefly outlined. As was noted earlier, progress during the stimulation phase is monitored by ultrasound and treatment may cease (be ‘cancelled') if poor or excessive response is anticipated. We treated the outcomes of these cycles as missing data to be imputed. This would appear to be appropriate for the purposes of investigating the dose-response of gonadotropin and effects of other predictors. The alternative would be to give these cycles a value of zero, which would not distinguish between under and over-responders. We made use of data relating to the intermediate measurements used to track progress (follicle counts on days 8 and 10 of the stimulation phase) and to the total dose of gonadotrophin delivered by the end of the period. While it would not be appropriate to adjust for these intermediate outcomes as covariates, they are likely to be strongly predictive of egg count, and are therefore useful for the purposes of imputation. A caveat is that the follicle counts were themselves subject to substantial amounts of missing data. We included these variables together with the covariables to be included in the analysis and a categorical variable reflecting patient ID in the multiple imputation model. We generated three completed datasets using the chained equations method implemented in the mi package in R (Su, et al., 2011). We conducted diagnostics, not displayed here, to investigate the suitability of the imputed values. Although we considered that the imputation model could probably be improved in some aspects, we did not pursue this due to the relatively low rate of missingness.

6.1.8. Why take a Bayesian approach?

Our primary motivation for adopting a Bayesian approach was pragmatism. Bayesian methods offer great flexibility, which we anticipated would facilitate the subsequent extension to multistage models.

6.1.9. Fitting the model

Four models were fitted to the completed datasets using MCMC in the software RStan (Stan Development Team, 2017). We fitted both lognormal and overdispersed poisson models, both with and without allowance for clustering of repeated measurements. Allowance for clustering was achieved by the addition of normal and gamma-distributed patient-level random effects, respectively. Vague prior distributions were used so that inferences would be based on the available data.

6.1.10. Model diagnostics

Diagnostics were performed for the overdispersed repeated-measures count model. Our approach was to consider diagnostic checking of the fitted model before making inferences from the regression parameters. Model checking was performed by conducting posterior predictive checks, similar to those described for complete cases above, and by constructing plots of binned residuals (Gelman et al., 2000). We calculated both 'realised' and 'replicated' residuals, where the former are calculated as the difference between observed egg counts and the fitted egg counts predicted by the model and the latter are calculated as the difference between egg counts from hypothetical replicate datasets generated from the posterior predictive distribution and the model predictions. The latter were used to construct 95% predictive intervals.

Plots of egg counts and of relevant test statistics from the posterior predictive distribution did not reveal substantial inconsistencies between the model and data. There is a tendency for datasets predicted by the model to contain excessive maximum values, although this appears to be consistent with chance (we could summarise the probability of the discrepancy using a one-sided Bayesian posterior p-value of 0.93, where values close to 1 or 0 would indicate a `statistically significant' discrepancy). The binned residual plots are displayed in Figure 13. If the model assumptions hold, the average residuals should be scattered around zero, and should not display any pattern. Although the residuals for the low to medium fitted values appear to be reasonably scattered about zero, there is a clear tendency for negative residuals in the higher bins, suggesting that for patients expected to have high egg counts, the model predicts even higher counts. However, even for the patients with the higher expected values, the size of the overestimation is typically quite small, amounting to an average residual of around -2 or smaller. We considered the model to be reasonable for the purposes of inference on the basis of these checks.

6.1.11. Comparison of fitted models

Estimates and 95% credible intervals (CIs) from the fitted models are displayed in Table 5 and Table 6. These have been exponentiated for ease of interpretation. It is apparent from the tables that estimates from all four models are similar. It should also be noted that the interpretation of parameters from the lognormal and Poisson models are similar. We would therefore not reach substantively different conclusions in relation to any of the parameters under the different models.

6.1.12. Discussion of models for stimulation response

Having investigated two plausible candidate distributions for egg yield, we opted to model stimulation response using count models in our subsequent work. While we reached similar conclusions regarding effects on the mean response in the present example, posterior predictive checks also suggested some discrepancies between the lognormal model and the observed egg yields. In particular, overprediction of the mean, variance and maximum yields could lead to a surfeit of erroneous predictions. Given that the range of egg yields which can be considered both sufficient and safe is relatively narrow, this could have material implications for the clinical utility of the model. Accordingly, we use count models to represent stimulation response subsequently.

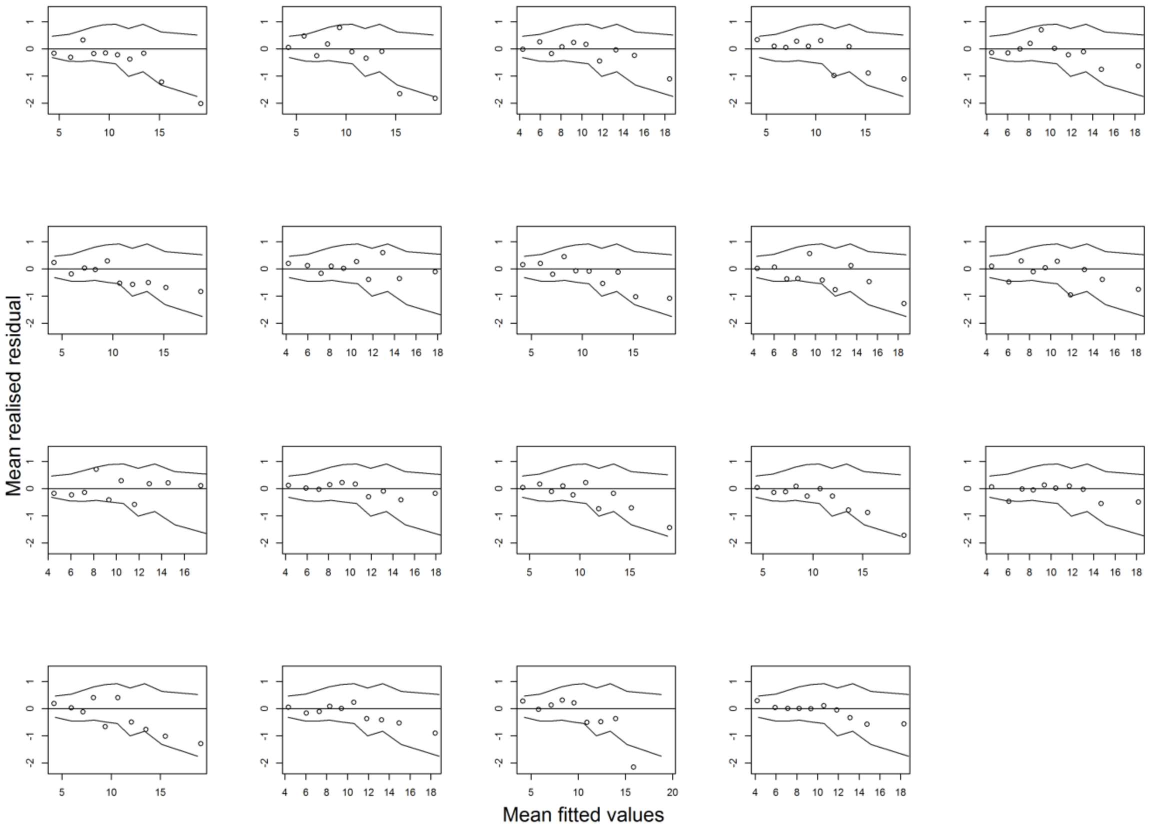


Figure 13: Plots of average realised residuals (y-axis) against average fitted values (x-axis) within bins for 19 draws of the posterior distribution of the overdispersed multilevel Poisson model coefficients. Grey lines are 95% predictive bounds.

Parameter	**Overdispersed repeated- measures Poisson**	**Overdispersed Poisson: ignore clustering**	**Log-normal w/o clustering**	**Log-normal repeated-measures model**
Intercept	8.91 (7.79 to 10.22)	9.08 (7.94 to 10.37)	7.30 (6.08 to 8.73)	7.33 (6.22 to 8.64)
LDR 75-150 IU	Ref	Ref	Ref	Ref
Antagonist 75-150 IU	0.76 (0.67 to 0.86)	0.77 (0.68 to 0.87)	0.72 (0.61 to 0.85)	0.71 (0.60 to 0.83)
LDR 187-250 IU	1.12 (1.01 to 1.25)	1.11 (1.00 to 1.24)	1.12 (0.96 to 1.31)	1.12 (0.97 to 1.30)
Antagonist 187 – 250 IU	1.08 (0.90 to 1.30)	1.07 (0.89 to 1.28)	1.12 (0.87 to 1.45)	1.12 (0.87 to 1.44)
LDR 300 IU	1.17 (1.03 to 1.33)	1.14 (1.01 to 1.30)	1.17 (0.99 to 1.40)	1.18 (1.00 to 1.41)
Antagonist 300 IU	1.04 (0.91 to 1.18)	1.02 (0.90 to 1.15)	1.06 (0.90 to 1.26)	1.07 (0.91 to 1.25)
LDR 375 IU	1.18 (0.92 to 1.51)	1.11 (0.87 to 1.40)	1.14 (0.82 to 1.56)	1.18 (0.86 to 1.62)
Antagonist 375 IU	1.11 (0.90 to 1.37)	1.09 (0.88 to 1.34)	1.16 (0.89 to 1.50)	1.14 (0.88 to 1.47)
LDR 450 IU	1.07 (0.87 to 1.33)	1.01 (0.82 to 1.24)	0.99 (0.76 to 1.33)	1.04 (0.80 to 1.35)
Antagonist 450 IU	0.94 (0.76 to 1.17)	0.90 (0.73 to 1.12)	0.85 (0.65 to 1.12)	0.89 (0.67 to 1.18)
Usor operator: A	Ref	Ref	Ref	Ref
B	0.98 (0.91 to 1.04)	0.97 (0.91 to 1.04)	1.01 (0.92 to 1.11)	1.02 (0.93 to 1.12)
C	1.04 (0.94 to 1.16)	1.04 (0.93 to 1.16)	1.12 (0.96 to 1.31)	1.11 (0.95 to 1.31)
D	0.68 (0.51 to 0.89)	0.70 (0.53 to 0.93)	0.72 (0.52 to 1.03)	0.72 (0.52 to 1.01)
E	0.78 (0.71 to 0.86)	0.78 (0.71 to 0.86)	0.72 (0.63 to 0.82)	0.72 (0.64 to 0.82)
F	0.86 (0.78 to 0.97)	0.86 (0.77 to 0.96)	0.85 (0.73 to 0.99)	0.85 (0.73 to 0.99)
G	0.95 (0.87 to 1.05)	0.95 (0.86 to 1.04)	0.95 (0.82 to 1.09)	0.95 (0.83 to 1.10)
H	0.93 (0.84 to 1.02)	0.93 (0.84 to 1.02)	0.96 (0.85 to 1.09)	0.97 (0.85 to 1.10)
I	0.77 (0.70 to 0.84)	0.77 (0.70 to 0.84)	0.76 (0.67 to 0.86)	0.76 (0.67 to 0.87)
J	0.70 (0.56 to 0.88)	0.69 (0.54 to 0.88)	0.46 (0.33 to 0.63)	0.46 (0.33 to 0.63)
Attempt No: 1st	Ref	Ref	Ref	Ref
2nd	1.05 (0.99 to 1.11)	1.05 (0.99 to 1.11)	1.11 (1.02 to 1.20)	1.10 (1.02 to 1.19)
3rd or 4th	1.19 (1.07 to 1.32)	1.18 (1.07 to 1.31)	1.32 (1.14 to 1.52)	1.32 (1.14 to 1.52)

Table 5: Posterior means and 95% credible intervals for exponentiated parameter estimates from fitted models (rate ratios for Poisson models, multiplicative effects for lognormal models).

Parameter	Overdispersed repeated-measures Poisson	Overdispersed Poisson w/o clustering	Log-normal w/o clustering	Log-normal repeated measures
Antral follicle count: < 10	Ref	Ref	Ref	Ref
11 to 16	1.16	1.16	1.20	1.20
	(1.11 to 1.23)	(1.10 to 1.23)	(1.11 to 1.30)	(1.11 to 1.30)
16 to 52	1.29	1.29	1.34	1.34
	(1.20 to 1.38)	(1.20 to 1.38)	(1.22 to 1.47)	(1.14 to 1.52)
Age (SDs)	0.87	0.87	0.86	0.86
	(0.85 to 0.89)	(0.85 to 0.90)	(0.83 to 0.89)	(0.83 to 0.89)
Age^2 (SDs)	0.96	0.96	0.96	0.96
	(0.94 to 0.99)	(0.94 to 0.99)	(0.94 to 0.99)	(0.94 to 0.99)
Log(AMH) (SDs)	1.35	1.33	1.41	1.42
	(1.30 to 1.40)	(1.28 to 1.38)	(1.34 to 1.48)	(1.35 to 1.50)
Gonadotrophin: HMG	Ref	Ref	Ref	Ref
rFSH	1.15	1.15	1.17	1.16
	(1.07 to 1.24)	(1.07 to 1.24)	(1.05 to 1.28)	(1.05 to 1.29)
Unexplained fertility	1.07	1.06	1.13	1.13
	(1.00 to 1.14)	(1.00 to 1.13)	(1.04 to 1.23)	(1.04 to 1.23)
Mild tubal	1.01	1.00	1.02	1.02
	(0.94 to 1.08)	(0.93 to 1.07)	(0.94 to 1.11)	(0.93 to 1.12)
Severe tubal	0.92	0.91	0.91	0.92
	(0.77 to 1.09)	(0.77 to 1.08)	(0.73 to 1.14)	(0.74 to 1.14)
Mild male factor	0.99	0.99	1.00	1.00
	(0.93 to 1.05)	(0.93 to 1.05)	(0.93 to 1.09)	(0.92 to 1.09)
Severe male factor	1.11	1.11	1.09	1.08
	(0.88 to 1.40)	(0.88 to 1.40)	(0.77 to 1.51)	(0.78 to 1.50)
Endometriosis	0.94	0.94	0.97	0.98
	(0.85 to 1.06)	(0.84 to 1.05)	(0.84 to 1.12)	(0.84 to 1.12)
Endometrioma	0.87	0.88	0.91	0.90
	(0.75 to 1.02)	(0.77 to 1.00)	(0.74 to 1.12)	(0.73 to 1.12)
Protocol: Old	Ref	Ref	Ref	Ref
New protocol (V1)	0.87	0.86	0.82	0.82
	(0.81 to 0.93)	(0.80 to 0.92)	(0.75 to 0.89)	(0.75 to 0.89)
New protocol (V2 & V3)	0.90	0.88	0.88	0.89
	(0.79 to 1.02)	(0.77 to 1.00)	(0.74 to 1.05)	(0.75 to 1.05)
New protocol (V4)	0.84	0.83	0.80	0.80
	(0.74 to 0.94)	(0.74 to 0.94)	(0.68 to 0.94)	(0.68 to 0.93)
BMI (SDs)	1.01	1.02	1.02	1.02
	(0.99 to 1.04)	(0.99 to 1.04)	(0.98 to 1.05)	(0.98 to 1.05)
Hyperparameters				
γ (rate, shape of patient-level random effect)	26.7	-	-	-
	(13.9 to 59.8)			
ζ (rate, shape of cycle-level random effect)	11.4	7.56	-	-
	(8.40 to 16.4)	(6.70 to 8.53)		
σ (level 1 SD)	-	-	0.65	0.60
			(0.64 to 0.66)	(0.58 to 0.62)
σ_b (level 2 SD)	-	-	-	0.25
				(0.21 to 0.29)

Table 6: Posterior means and 95% credible intervals for exponentiated parameter estimates from fitted models (rate ratios for Poisson models, multiplicative effects for lognormal models).

6.2 How to model responses at the embryo culture stage?

6.2.1. Motivation

Our review of outcome measures in IVF RCTs (Journal Article 2, Chapter 4) suggested that the quality of embryos arising from embryo culture is not reported in a standardised manner, with some authors reporting this in a binary fashion (answering the question 'is this embryo good quality – yes or no?') and others using esoteric scales (Journal Article 2). Previous guidelines for the evaluation of embryo morphology have suggested that three dimensions should be taken into consideration: cell number (representing size or growth), evenness and fragmentation (Cutting, et al., 2008). A validation exercise suggested that growth and fragmentation were strongly predictive of pregnancy (Stylianou 2012). In the following, we do not attempt to combine morphology parameters into a single index. Instead, we consider the triplet of morphology parameters as a multivariate response. This approach allows us to estimate different covariate effects for the different quality measures, as well as accommodating (and quantifying) associations between them. Importantly, approaches based on the joint modelling of responses of different types can then be extended to include other stages of treatment.

As for the stimulation stage, we attempted to answer a clinical question using real data, using this as a vehicle for methodological development. Whereas, for the stimulation stage, we wrote a manuscript for publication detailing our clinical findings, we did not attempt to do the same here. This is because we were not confident in the quality of the dataset used for this analysis. The following analyses of embryo quality parameters should not be interpreted as anything other than an exercise in method development. Our focus here is on model checking and comparison.

6.2.2. Introduction to the dataset and the clinical research question: a comparison of two incubators

A pseudo randomised comparison between two embryo incubators was conducted at St Mary's Hospital Department for Reproductive Medicine. Patients' embryos were allocated (in batches) to either a standard or experimental incubator on the basis of availability. In the

Variable	Summary
Attempt	
1st	530 (77%)
2nd	124 (18%)
3rd	26 (4%)
ICSI	374 (55%)
IVF	306 (45%)
Age (years)	32
	29 to 35
	21 to 42
Partner Age (years)	34
	31 to 39
	23 to 59
Transfer Day	
3	420 (62%)
5	260 (38%)
Number of oocytes	10
	7.75 to 13
	3 to 26
Incubator	
Embryoscope	339 (50%)
Hunter	279 (41%)
Split	62 (9%)

Table 7: Summary of cycles in the dataset. Five-number summary for continuous/numeric variables, frequency and percentage for categorical variables.

present example, we were interested in the effect of incubator on the three embryo morphology parameters described above: number of cells, evenness, and fragmentation. There are 4750 embryos in the dataset, from 680 cycles in 610 patients. The cycles were conducted between 2013 and 2014. 543 (89%) patients have 1 cycle, 64 (10%) have 2 cycles, 3 (0%) have 3. The median (IQR) number of embryos per cycle is 6 (4 to 9), range (1 to 42). Numerical summaries of the cycle characteristics and embryo outcomes are displayed in Table 7 and Table 8. 4605 embryos (97%) had no missing data, so we used complete cases to develop our embryo models.

Outcome	Day2	Day3
Cell Number	4	7
	3 to 4	6 to 8
	1 to 11	1 to 16
	0%	*1%*
Evenness	*2%*	*2%*
1	116 (2%)	122 (3%)
2	1186 (26%)	1445 (31%)
3	2390 (51%)	2561 (55%)
4	954 (21%)	534 (11%)
Fragmentation Degree	*2%*	*1%*
1	108 (2%)	119 (3%)
2	664 (14%)	872 (19%)
3	1404 (30%)	1617 (35%)
4	2470 (53%)	2052 (44%)
Fragmentation Pattern	*2%*	2%
A	840 (18%)	630 (14%)
B	2275 (60%)	2799 (60%)
C	15 (0%)	17 (0%)
D	375 (8%)	533 (11%)
E	568 (12%)	597 (13%)
F	73 (2%)	84 (2%)

Table 8: Summary of embryo-level outcomes in the dataset. Five-number summary for continuous/numeric variables, frequency and percentage for categorical variables.

6.2.3. A multivariate Normal model for embryo morphology parameters

In our initial embryo morphology model, we represented each of the three response variables using linear regression. In order to accommodate and quantify relationships between the response variables, we modelled them simultaneously by allowing the patient-level random effects to be correlated.

Mathematical representation

Let *j* = 1, … , *J* index treatment cycles in the dataset and *i = 1, …,* I_j index embryos nested within cycles. Here, we do not account for clustering of multiple cycles within patients, because only 67 (10%) have as many as 2 cycles. Let y_{ij}^{N} , $y_{ij}^{E}$and y_{ij}^{F} represent the outcomes cell number, cell evenness and degree of fragmentation for embryo *i* of cycle *j*, respectively. The outcomes y_{ij}^{N}, $y_{ij}^{E}$and y_{ij}^{F} are modelled as functions of covariates and random effects:

$$
\begin{aligned}
y_{ij}^{N} &= \boldsymbol{X}_{ij}\boldsymbol{\beta}_{N} + z_{j}^{N} \quad + e_{ij}^{N} \\
y_{ij}^{E} &= \boldsymbol{X}_{ij}\boldsymbol{\beta}_{E} + z_{j}^{E} \quad + e_{ij}^{E} \\
y_{ij}^{F} &= \boldsymbol{X}_{ij}\boldsymbol{\beta}_{F} + z_{j}^{F} \quad + e_{ij}^{F} \\
& e_{ij}^{N} \sim N(0, \sigma_{N}^{2}) \\
& e_{ij}^{E} \sim N(0, \sigma_{E}^{2}) \\
& e_{ij}^{F} \sim N(0, \sigma_{F}^{2})
\end{aligned}
$$

$$
\begin{bmatrix} z_{j}^{N} \\ z_{j}^{E} \\ z_{j}^{F} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_{N}^{2} & \rho_{1}\omega_{N}\omega_{E} & \rho_{2}\omega_{N}\omega_{F} \\ \rho_{1}\omega_{E}\omega_{N} & \omega_{E}^{2} & \rho_{3}\omega_{E}\omega_{F} \\ \rho_{2}\omega_{F}\omega_{N} & \rho_{3}\omega_{F}\omega_{E} & \omega_{F}^{2} \end{bmatrix} \right)
$$

The model contains three submodels corresponding to the three aforementioned morphology parameters. In the submodels, $\boldsymbol{X}_{ij}$ is a 1 x p row-vector of predictor variables and $\boldsymbol{\beta}_{N}$, $\boldsymbol{\beta}_{E}$ and $\boldsymbol{\beta}_{F}$ are p x 1 vectors of fixed regression coefficients for these predictors. In this case, the same covariate vector x_{ij} appears in all three submodels, but this is not required in general. Different covariates and different representations of covariates (transformations, interactions) may appear in the different parts of the model. Different regression parameters are estimated in each submodel so that the relationships between each morphology parameter and covariates are permitted to differ. z_{j}^{N} , z_{j}^{E} and z_{j}^{F} are cycle-level random scalars. These account for the correlation between measurements of each morphology parameter corresponding to embryos created in the same cycle. A multivariate Gaussian distribution is specified for these level 2 residuals. An unstructured correlation structure is used, to permit the estimation of associations between the variables.

6.2.4. Model fitting

The joint model was fitted using the MCMC software Stan (Stan Development Team, 2017). In order to fit the models, it is necessary to specify priors on the model parameters and hyperparameters. Normal (0, 1000^2) priors are placed on each of the regression coefficients. Default improper priors over non-negative real numbers are used for each of the level 1 and level 2 residual standard deviations. An LKJ Correlation (0.4) prior was placed on the correlation matrix (Stan Development Team, 2017). This is an informative prior. Specifically, it represents a prior belief that the correlation matrix is unlikely to be represented by the identity matrix; or, in other words, that the morphology parameters are unlikely to be uncorrelated.

Three Markov chains were each run for 1000 iterations. The first 500 from each were discarded as burn in. The remaining 500 from each were pooled to provide 1500 draws from the posterior distribution. Convergence was assessed using traceplots and the Gelman-Rubin statistic (Gelman and Rubin, 1992).

6.2.5. Model checking for the multivariate Normal model

We have modelled three discrete outcomes using a model for continuous distributions. Before using the model to make inferences about these morphology parameters, its suitability for this purpose should be assessed. As for our investigation of the stimulation stage, we use the posterior predictive distribution for this purpose. Briefly, this involves simulating new datasets representing hypothetical replications of the observed outcomes, and comparing the characteristics of the simulated data to those of the actual data. Systematic discrepancies may highlight a need for model improvement.

Plots of predictions against observed data

Figure 14 shows histograms of the observed cell numbers together with the cell numbers from 19 randomly selected replicated datasets. It is immediately apparent that the replicates display considerably greater symmetry than the observed counts, albeit in a similar range and with a similar mean. The large spike at 4 is clearly not captured by the model. This suggests that alternative models (for example, using log (cell number) might be worth considering.

The morphology parameters are bounded; cell number has a minimum possible value of zero and evenness and fragmentation degree are 1 to 4 ordinal scales. The model, by contrast, does not result in a bounded predictive distribution. The extent to which the model yields predictions outside of this range therefore warrants inspection.

We calculated the proportion of replicated datasets containing predictions lying outside of the range dictated by the measurement scale. Seventy-one per cent of the replicated datasets included predictions of cell numbers below zero, and all of them included values less than 1 for evenness and fragmentation degree. However, the actual proportions of observations falling below the lower bound were very low. For cell number, the mean proportion (rescaled to %) was 0.03%, with a maximum proportion of 0.13%. For evenness the mean was 0.57%, with a maximum of 0.98%. For fragmentation degree, the mean was 0.26%, with a maximum of 0.63%. We would almost certainly consider this level of error to be acceptable. The situation was worse at the upper end of the distribution however. None of the replicated datasets contained cell numbers as great as

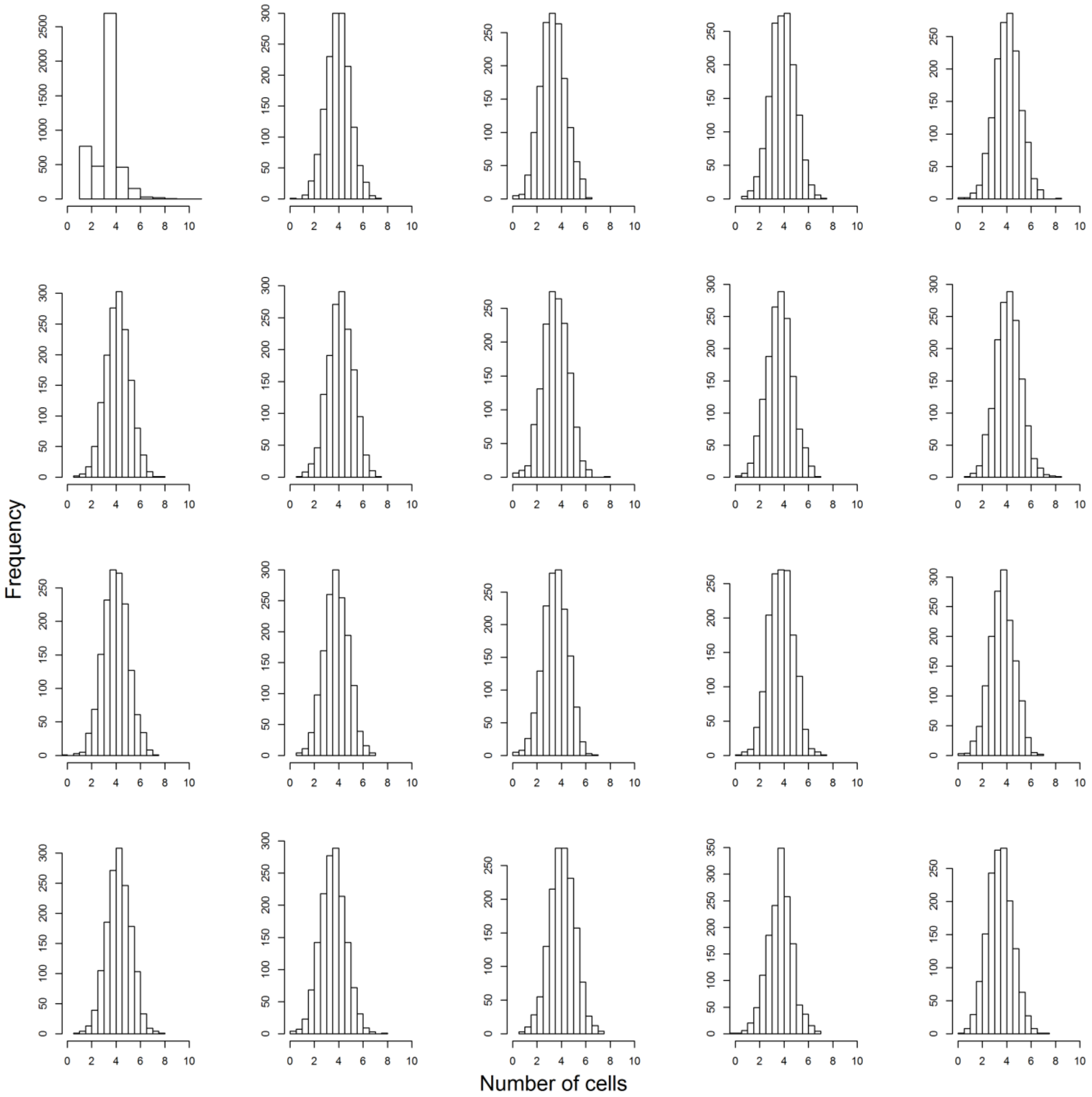


Figure 14: Observed day 2 cell counts (top left) and hypothetical replications drawn from the posterior predictive distribution of the multivariate Normal model (note differences in y-axes).

11, the greatest observed in the source data, although it is worth noting that only 5(0.1%) observations in the observed data were greater than 8. Only 21% contained predictions greater than 8, and no replicate had more than 0.07% of predictions exceeding this. They all contained values of fragmentation degree and evenness greater than 4. Nontrivial amounts of fragmentation and evenness predictions in each set exceeded 4; the mean and minimum proportions were 6.8% and 5.4% for evenness and 21.5% and 18.9% for fragmentation. Moreover, while the average (max) proportion of predictions exceeding 5 was 0.21% (0.56%) for evenness, the average (min) for fragmentation degree was 1.7% (1%).

Figure 15 shows the means and standard deviations from the predicted datasets together with the corresponding values from the observed data. The model captures the means and SDs of all three parameters well.

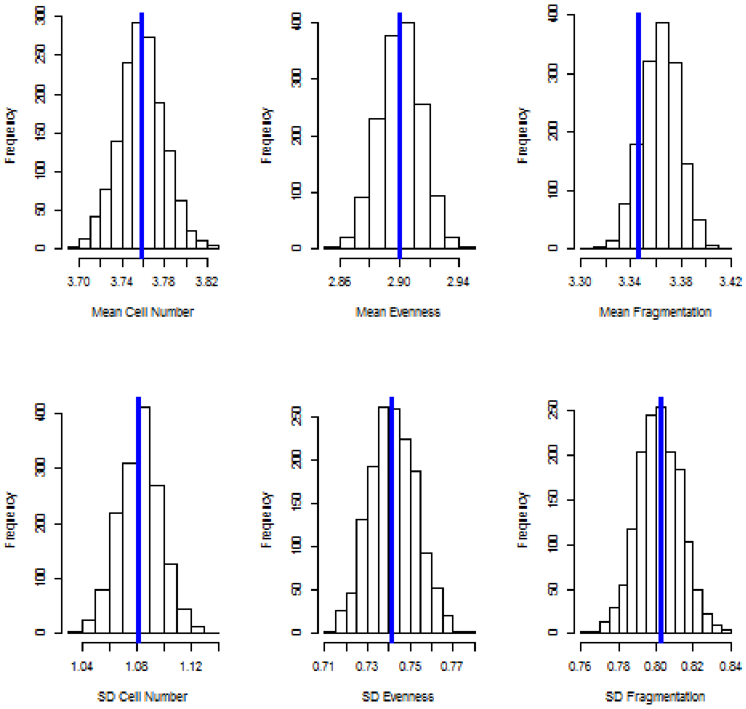


Figure 15: Histograms of the means (top row) and standard deviations (bottom row) from the replicated datasets drawn from the multivariate Normal model, for cell number, evenness and fragmentation degree. Blue lines are observed values from the dataset.

Residuals from the fitted model

Further graphical assessment of the model was conducted using the conditional residuals. As we did when evaluating the fit of our stimulation models (see section 6.1.10), we computed both realised and replicated residuals (Gelman et al., 1999). As the data are discrete, we plotted averaged residuals against averaged fitted values within bins (Gelman et al., 1999). These are displayed for each morphology parameter in Figure 16, Figure 17, and Figure 18.

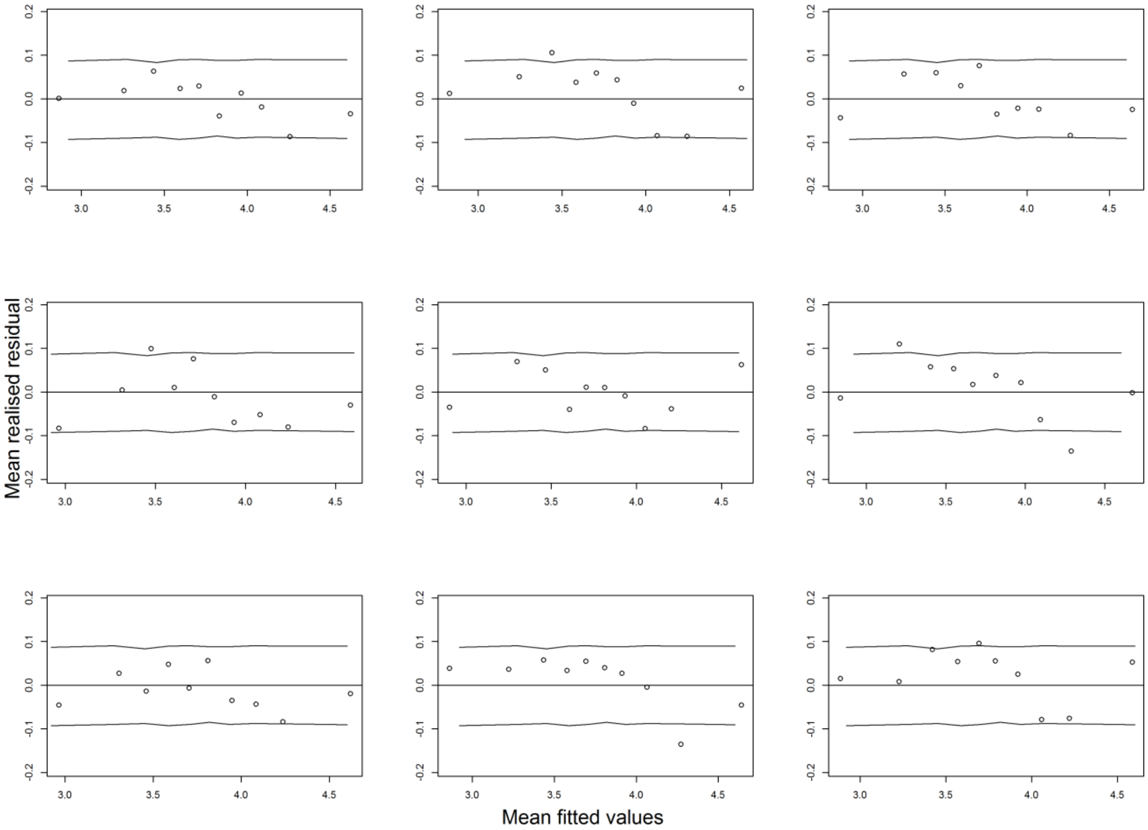


Figure 16: Averaged realised residuals plotted against averaged fitted values for day 2 cell number, for 9 randomly selected draws from the posterior distribution of the multivariate Normal model. Grey lines show 95% bounds of the distribution of averaged replicated residuals. The horizontal line indicates an averaged residual value of zero.

The plots for cell number and evenness suggest that the model is reasonable for these parameters; averages of realised residuals are close to zero and are generally consistent with the replicated residuals. However, the plots for fragmentation degree show systematic error, with underestimation for lower values and overestimation at higher values. The inconsistency between the observed and fitted values is thrown into relief by the discrepancy between the realised and replicated residuals.

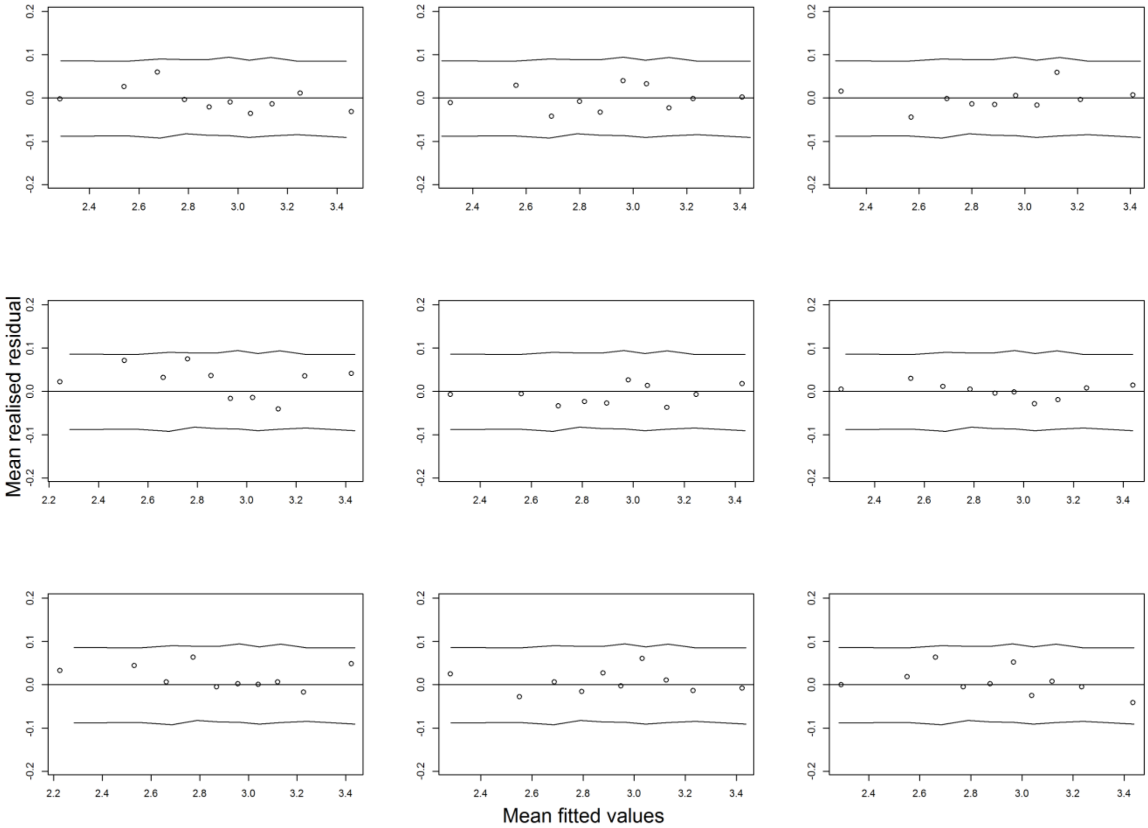


Figure 17: Averaged realised residuals plotted against averaged fitted values for day 2 cell evenness, for 9 randomly selected draws from the posterior distribution of the multivariate Normal model. Grey lines show 95% bounds of the distribution of averaged replicated residuals. The horizontal line indicates an averaged residual value of zero.

6.2.6. **Inference from the multivariate Normal model**

The model checks suggest that there may be scope for model improvement, particularly in relation to cell fragmentation degree. However, we present the inference from the model here. From a methodological perspective, there is interest in whether different approaches yield different conclusions. Posterior means of the effects of the conventional versus experimental incubator (95% credible intervals) were -0.15 (-0.22 to -0.07) for cell number, 0.09 (0.04 to 0.15) for cell evenness, and 0.11 (0.04 to 0.17) for cell evenness. These suggest that there is no practical difference between the incubators in terms of morphology.

We also obtain estimates of the latent correlation between the response variables (Table 9). We will consider the question of how these can be interpreted in more detail in Chapter 7. For now we simply note the direction and size of the estimates. The correlation between cell number and cell evenness is estimated to be very small. The correlation between cell number and fragmentation degree is estimated as positive, but is at most moderate. There is a strong positive correlation between evenness and fragmentation degree.

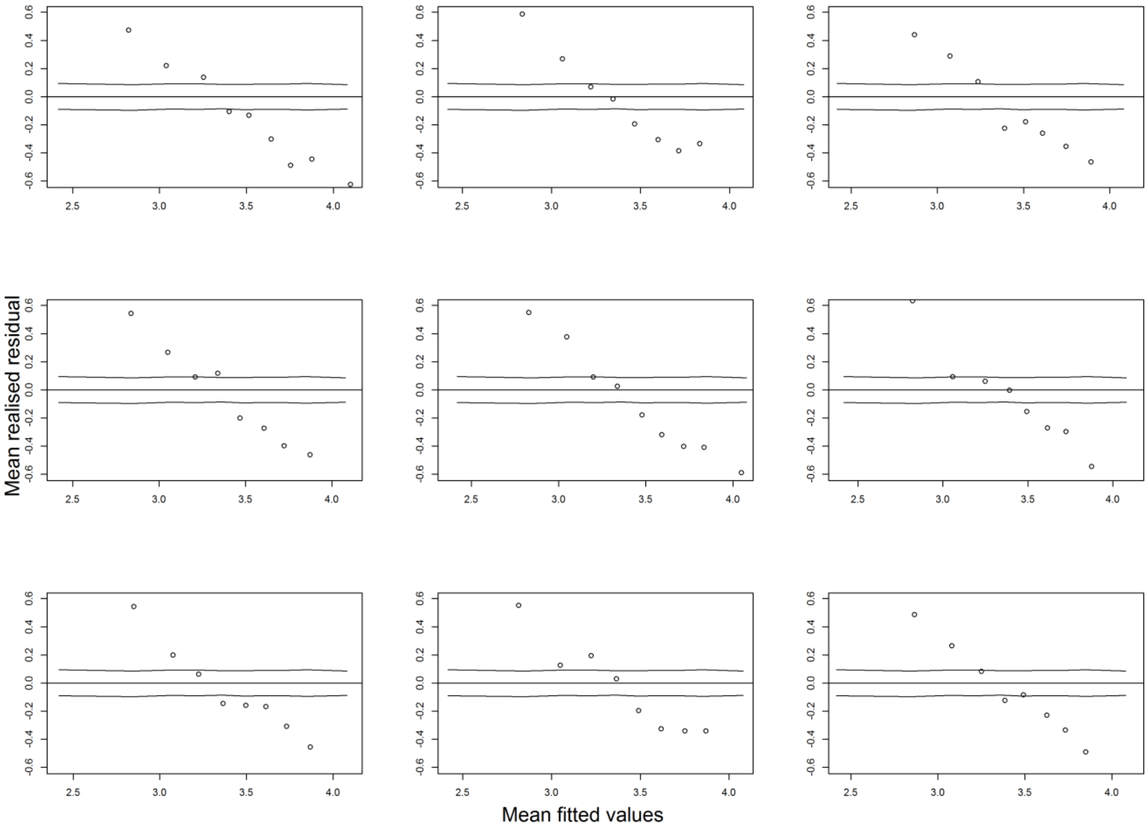


Figure 18: Averaged realised residuals plotted against averaged fitted values for day 2 cell fragmentation degree, for 9 randomly selected draws from the posterior distribution of the multivariate Normal model. Grey lines show 95% bounds of the distribution of averaged replicated residuals. The red line indicates an averaged residual value of zero.

	Cell Number	Cell Evenness	Fragmentation Degree
Cell Number	1.00	-0.01 (-0.13 to 0.12)	0.16 (0.05 to 0.27)
Cell Evenness	-0.01 (-0.13 to 0.12)	1.00	0.84 (0.78 to 0.89)
Fragmentation Degree	0.16 (0.05 to 0.27)	0.84 (0.78 to 0.89)	1.00

Table 9: Correlation matrix estimated from the joint Normal model. Posterior means and 95% credible intervals are displayed. Outcomes measured on day 2.

6.2.7. A mixed outcome model for embryo morphology parameters

On the basis of our model checks, there is likely to be scope for improvement of the multivariate Normal model. Here, we consider alternative representations of the morphology parameters. The observed cell numbers are the product of an underlying continuous growth process, whereby cells divide at some rate. We therefore apply a log base 2 transformation for cell number, and use a linear submodel. For the ordinal outcomes evenness and fragmentation degree, the linear models made predictions outside of the scale. We use cumulative logit models here in order to avoid this undesirable feature.

Once again, we connect the submodels by specifying a multivariate Normal distribution for the level 2 random effects. Accordingly, this model serves as our first foray into joint modelling of mixed response types, which we envision will be extensible for the purposes of multistage modelling. We proceed with a mathematical representation of our embryonic joint model. For simplicity, we recycle some of the sub- and superscripts appearing in our presentation of the multivariate Normal model in 6.2.3. This should not be taken to indicate equivalence between the parameters in the two models.

Mathematical representation

For embryo *i* nested in cycle *j* we model y_{ij}^{N}, $\log_2$ (cell number), using a linear regression model:

$$y_{ij}^{N} = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{N} + z_{j}^{N} + e_{ij}^{N}$$

$$e_{ij}^{N} \sim N(0, \sigma_{N}^{2})$$

We use cumulative logit models for the ordinal outcomes. For k = 1,2,3:

$$\text{logit}\left(\gamma_{kij}^{E}\right) = \alpha_{kE} - \boldsymbol{X}_{ij}\boldsymbol{\beta}_{kE} - z_j^E$$

$$\text{logit}(\gamma_{kij}^{F}) = \alpha_{kF} - \boldsymbol{X}_{ij}\boldsymbol{\beta}_{kF} - z_j^F$$

$$\begin{bmatrix} z_j^N \\ z_j^E \\ z_j^F \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_N^2 & o_1\omega_N\omega_E & o_2\omega_N\omega_F \\ o_1\omega_E\omega_N & \omega_E^2 & o_3\omega_E\omega_F \\ o_2\omega_F\omega_N & o_3\omega_F\omega_E & \omega_F^2 \end{bmatrix}\right)$$

Where $\boldsymbol{X}_{ij}$ is a row-vector of regressors, $\boldsymbol{\beta}_N, \boldsymbol{\beta}_{kE}$ and $\boldsymbol{\beta}_{kF}$ are each q x 1 vectors of regression coefficients, and z_j^N , z_j^E and z_j^F , are correlated cycle-level random effects. γ_{kij}^{E} and γ_{kij}^{F} are cumulative probabilities of embryo *i* in cycle *j* having a grade of *k* or lower for evenness and fragmentation degree respectively and α_{kE} and a_{KF} $(k = 1,2,3)$ are threshold parameters, corresponding to the log-odds of the embryo having this or a lower grade. If we set $\boldsymbol{\beta}_{1E} = \boldsymbol{\beta}_{2E} = \boldsymbol{\beta}_{3E}$ (and, similarly, $\boldsymbol{\beta}_{1F} = \boldsymbol{\beta}_{2F} = \boldsymbol{\beta}_{3F}$) then the model is subject to the *proportional odds* assumption, which states that covariate effects are constant across gradings. We started out by modelling the data with this assumption in place, although exploratory analyses suggested that this was probably a stretch (see S Table 16 and S Table 17, displaying ratios of cumulative probabilities for different levels of the model covariates).

6.2.8. Model checking for the mixed outcome model

Cell number

In the present version of the model, we log transformed cell number using a base value of 2. This was done to reflect the underlying process of cells doubling as an embryo grows, rather than due to any statistical concerns with the model of untransformed cell counts. Figure 9 shows the observed cell numbers together with those from 19 random draws from the posterior predictive. Figure 20 shows the means, SDs and maximum values from the replicated datasets together with the corresponding summary statistics from the observed data.

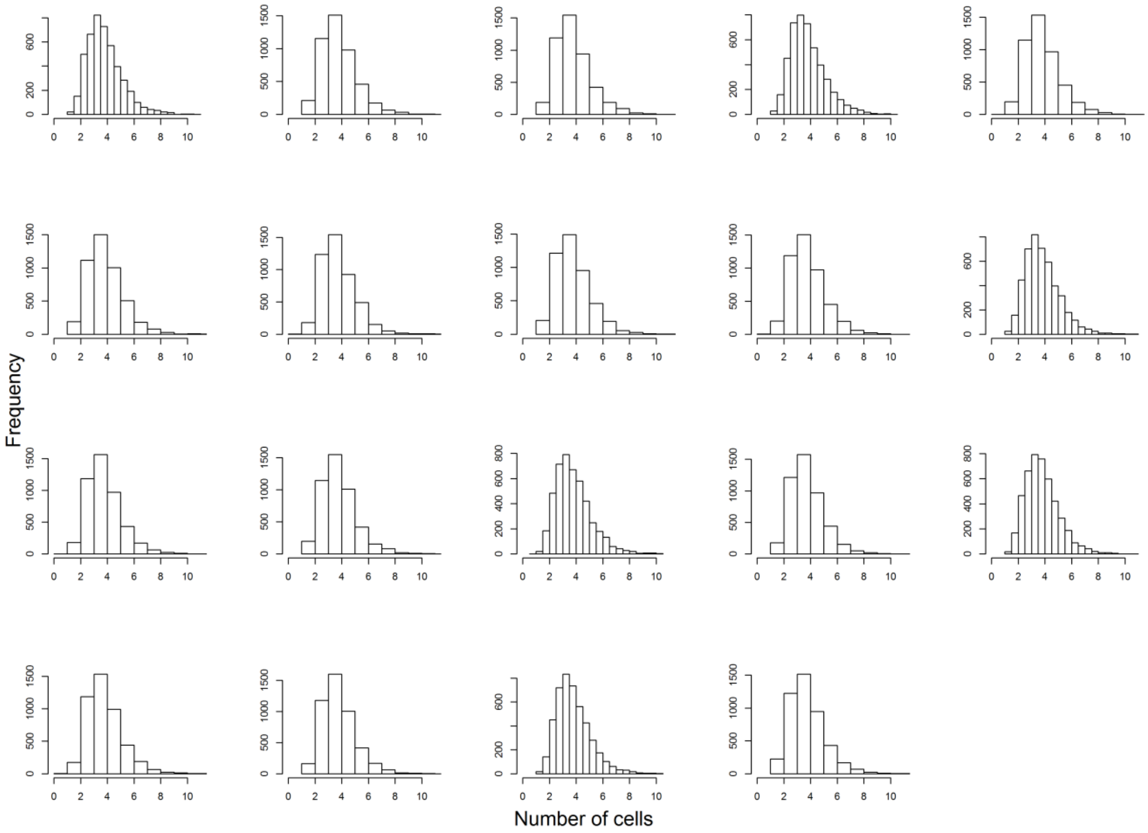


Figure 19: Observed cell counts (top left) together with predicted cell counts from 19 random draws from the posterior predictive distribution of the mixed outcome joint model. Note differences in y-axes.

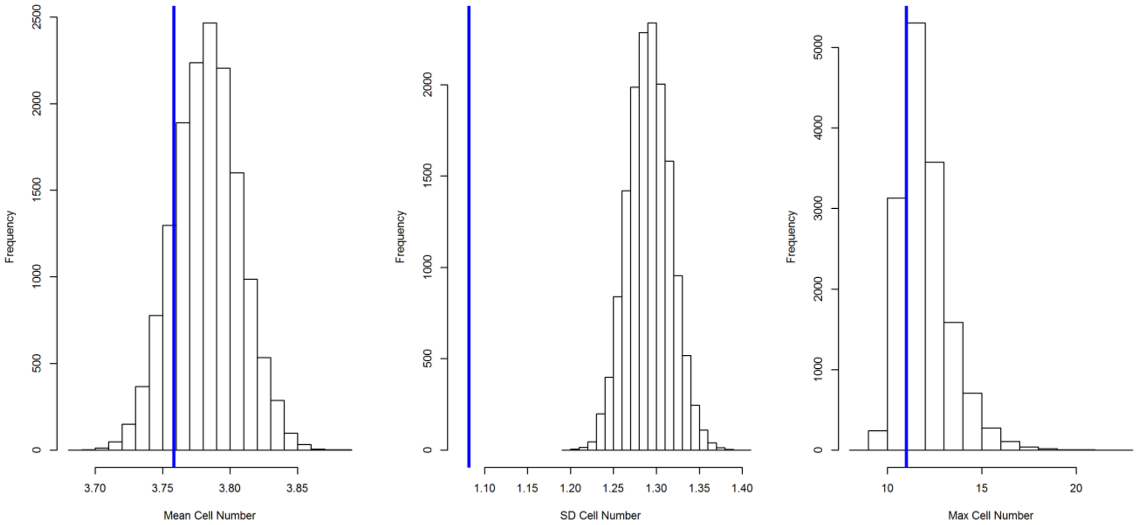


Figure 20: Histograms of the means (first plot), standard deviations (second plot) and maximum values (third plot) from replicated datasets, for cell number, drawn from the mixed outcome joint model, on the raw scale. Blue lines are observed values from the dataset.

The posterior replications here much more closely resemble the data than did those from the multivariate Normal model (Figure 19). This model to some extent captures the skew in the data, although it does not reproduce the large spike at a cell count of 4. An analysis of a future dataset might incorporate this observation by using a prior distribution that places a lot of probability mass at 4. The mean cell count is reasonably well captured by the model; most of the replicated datasets have values that are too high, but only by a negligible amount (Figure 20). The other two plots are perhaps more worrying. The variance is consistently overestimated (100% of the time), which may lead to unnecessarily imprecise predictions of cell number. The maximum values in the replications are also usually too high, although we would consider this to be consistent with chance by conventional standards (Bayesian one-tailed P = 0.78, where P > 0.95 would indicate incompatibility between model and data). In any case, we might not be too concerned if the model makes poor predictions at the extreme top end of the distribution, as long as it makes good predictions most of the time. It might be more relevant to look at quantiles at the upper end of the distribution. Figure 21 shows the observed and predicted 75th and 95th quantiles. The plots show that the predicted quantiles are systematically too high compared to the observed quantiles, although not by much. We are left with the impression that although the predictions are too diffuse compared to the data, that we would expect to see reasonable predictions for most embryos. For completeness, we also consider plots of averaged residuals against averaged fitted values on the raw scale from the cell number submodel (Figure 22). The plots suggest that the model systematically underestimates the cell number for lower values, by a negligible amount (on average). There also appears to be some relatively minor inconsistency between the 95% bounds of the distribution of replicated residuals and the realised residuals, with points falling outside the bounds in all of these plots.

Fragmentation and evenness

The multivariate normal model frequently made predictions outside the possible range of values for the ordinal responses evenness and fragmentation. This is a consequence of treating ordinal responses as continuous. The present mixed response model cannot make predictions outside the possible range for the ordinal outcomes, since we use cumulative logit submodels to predict the multinomial probabilities of an embryo having each grade. We can use the posterior predictive distribution to compare the distribution of grades under our model to the observed distribution. Figure 23 and Figure 24 display predicted and observed grades for cell evenness and fragmentation degree, for 9 draws from the posterior predictive distribution. The plots suggest that the model is a good fit to the data – the predicted grades closely resemble the observed grades for both ordinal outcomes.

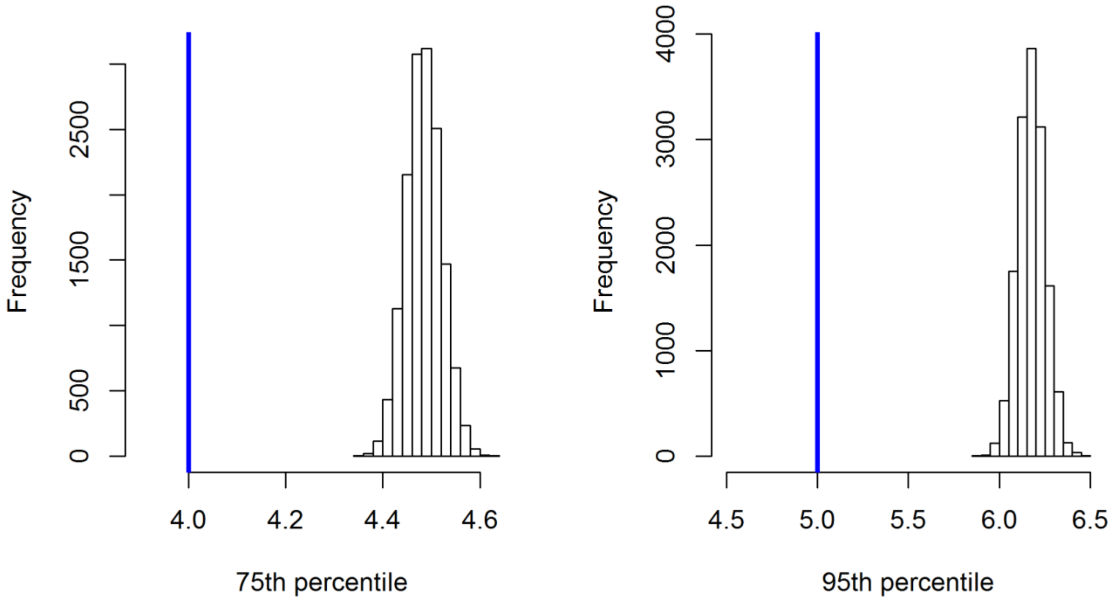


Figure 21: Histograms of the 75th (first plot), and 95th (second plot) percentiles from the replicated datasets drawn from the mixed outcome joint model, for cell number, on the raw scale. Blue lines are observed values from the dataset.

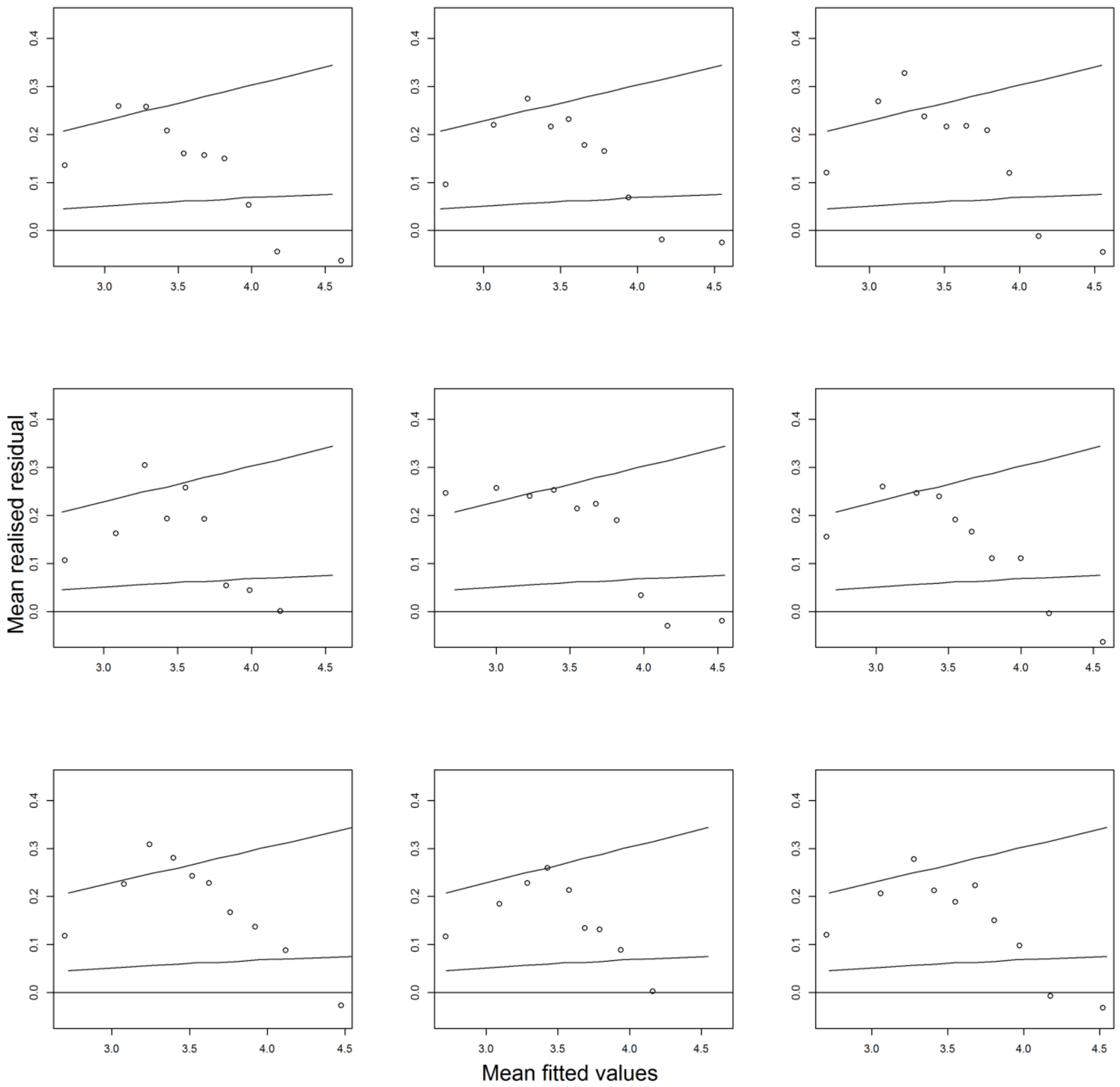


Figure 22: Averaged realised residuals plotted against averaged fitted values for day 2 cell fragmentation degree, for 9 randomly selected draws from the posterior distribution of the mixed outcome joint model. Grey lines show 95% bounds of the distribution of averaged replicated residuals. The horizontal line indicates an averaged residual value of zero.

6.2.9. **Inference from the mixed response model**

Exponentiated parameters (95% CIs) corresponding to the estimated multiplicative effect of the experimental compared to the conventional incubator are 1.05 (1.02 to 1.08), 0.78 (0.68 to 0.91) and 0.70 (0.58 to 0.87) for cell number, evenness and fragmentation, respectively. These are means and quantiles from the posterior distribution, and the second and third are odds ratios giving the relative odds of an embryos being in a higher rather than a lower grade. Accordingly, the estimates suggest an advantage of the experimental incubator in relation to growth rate (cell number) and a disadvantage in relation to evenness and fragmentation. Our conclusions are therefore consistent with the multivariate Normal model.

We also consider the estimated latent correlation matrix from the mixed response model (Table 10). Reassuringly, the correlation matrices arising from the two approaches are similar. Again, we see a strong positive correlation between cell evenness and fragmentation degree and a small to moderate positive correlation between cell number and fragmentation degree. The level 2 residual SDs in all submodels are rather large. A cycle with random effects one standard deviation above the mean would be expected to have embryos with 1.16 times the number of cells, 2.8 times the odds of a higher cell evenness grade and 1.68 times the odds of a higher fragmentation degree grade. Perhaps this is unsurprising; given the fact that the model predictors are estimated to have relatively little bearing on the morphology parameters, there remains a lot of variation to be explained. This is captured by a diffuse random effects distribution.

	Cell Number	Cell Evenness	Fragmentation Degree
Cell Number	1.00	0.07 (-0.05 to 0.19)	0.19 (0.09 to 0.30)
Cell Evenness	0.07 (-0.05 to 0.19)	1.00	0.80 (0.74 to 0.85)
Fragmentation Degree	0.19 (0.09 to 0.30)	0.80 (0.74 to 0.85)	1.00

Table 10: Correlation matrix estimated from the mixed response joint model. Posterior means and 95% credible intervals are displayed. Outcomes measured on day 2.

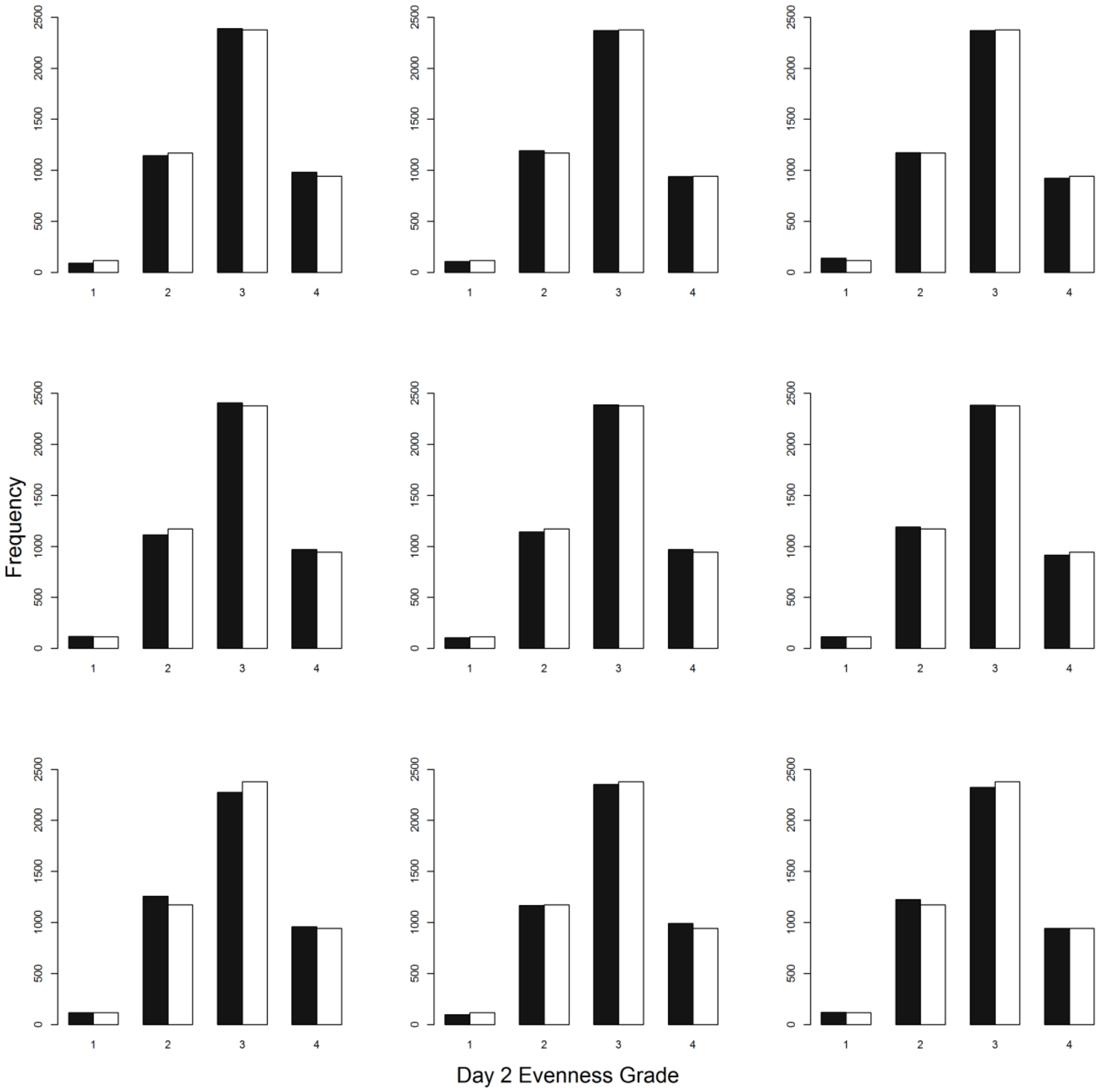


Figure 23: Distribution of predicted (black bars) and observed (white bars) evenness grades. Predictions are made from 9 draws from the posterior predictive distribution of the mixed outcome joint model .

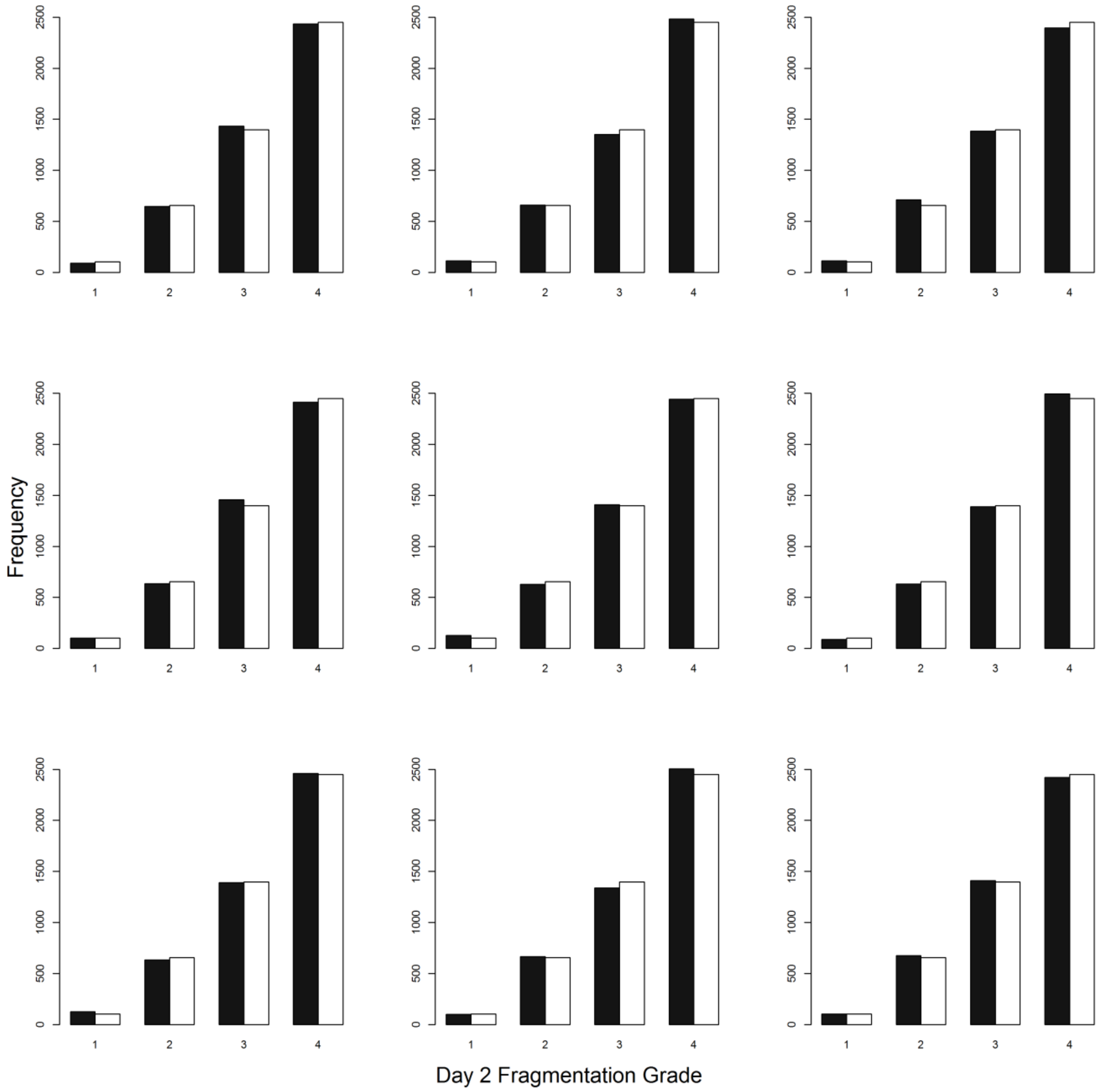


Figure 24: Distribution of predicted (black bars) and observed (white bars) fragmentation degree grades. Predictions are made from 9 draws from the posterior predictive distribution of the mixed outcome joint model.

6.2.10. Relaxing the proportional odds assumption in the mixed response model

The mixed response model used so far is subject to the proportional odds assumption, which corresponds to the assumption that covariate effects are constant across the levels of the ordinal grading scales. We conducted a further analysis where we relaxed this assumption (so that it was no longer the case that $\boldsymbol{\beta}_{1E} = \boldsymbol{\beta}_{2E} = \boldsymbol{\beta}_{3E}$, nor that $\boldsymbol{\beta}_{1F} = \boldsymbol{\beta}_{2F} = \boldsymbol{\beta}_{3F}$). This approach yields estimates of the effect of the experimental compared to the conventional incubator on the likelihood of getting a higher grade at each level on the scale. Odds ratios (95%) corresponding to the effect of the experimental compared to the conventional incubator on getting a higher evenness grade were 0.89 (0.60 to 1.39), 0.93 (0.78 to 1.12), and 0.61 (0.50 to 0.75) at grades 1,2 and 3 respectively. For fragmentation grade, the values were 0.74 (0.46 to 1.25), 0.74 (0.58 to 0.97) and 0.68 (0.54 to 0.85). Both the estimates for cell number and of the correlation between morphology parameters are completely unchanged compared to the proportional odds model. The estimates at grade k=3 for the ordinal scales suggest that the experimental incubator considerably decreases the chance of top grade embryos compared to the conventional incubator, which could not be discerned using the proportional odds model. To understand the practical relevance of these effects we created graphs comparing the impact of incubator to other sources of variation, such as age (Figure 25, Figure 26).

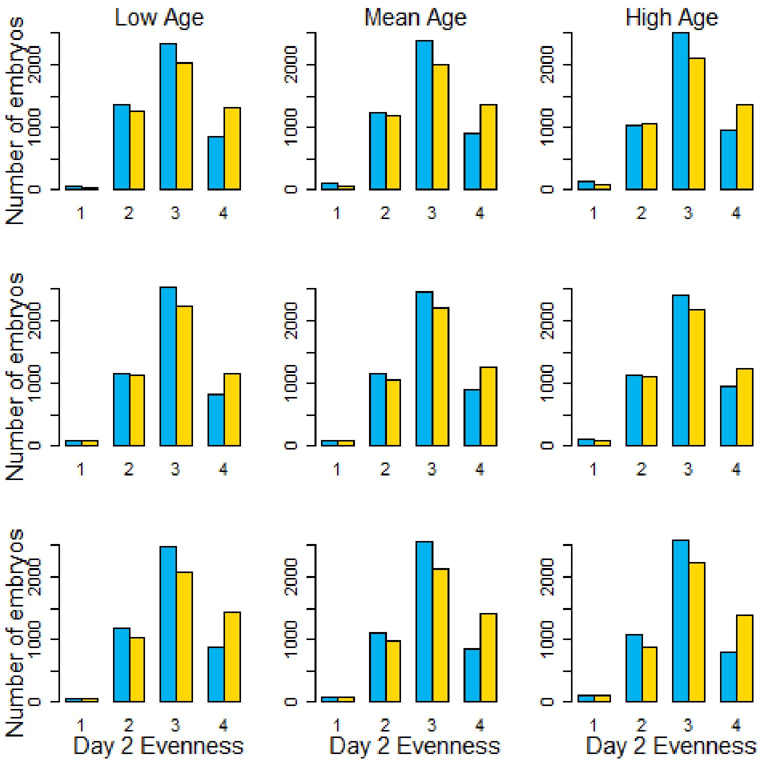


Figure 25: Distribution of day 2 cell evenness grades under 3 draws from the posterior predictive distribution of the mixed response joint model with non-proportional odds (rows), for a patient undergoing their first IVF attempt with a partner of mean age. Blue bars show response distribution under the experimental and gold show responses under the conventional incubator. Predicted responses are shown for low (-1 SD), medium (mean) and high (+1 SD) ages.

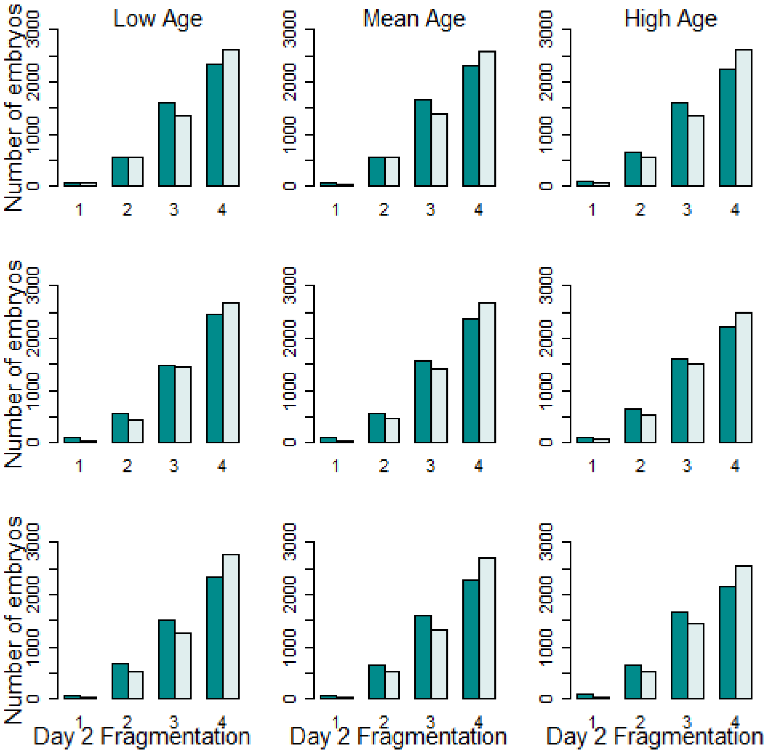


Figure 26: Distribution of day 2 cell fragmentation grades under 3 draws from the posterior predictive distribution of the mixed response joint model with non-proportional odds (rows) for a patient undergoing their first IVF attempt with a partner of mean age. Green bars show response distribution under the experimental and azure show responses under the conventional incubator. Predicted responses are shown for low (-1 SD), medium (mean) and high (+1 SD) ages.

Figure 25 suggests that the effects of incubator on cell evenness is substantial compared to age effects, with a greater number of grade 4 embryos obtained using the conventional incubator. By contrast, Figure 26 shows that effects on fragmentation are less pronounced, but certainly not trivial.

6.2.11. Discussion of embryo culture models

We considered three joint modelling approaches for the analysis of embryo responses, by way of an evaluation of the comparative efficacy of two incubators. Similar conclusions arise using all approaches; the conventional incubator performs similarly in terms of cell numbers, and improves evenness and fragmentation degree outcomes. Advantages of the mixed response models compared to the multivariate Normal include the fact that evenness and fragmentation degree are modelled as discrete variables, so that predictions arising from the model are both meaningful and interpretable. By relaxing the proportional odds assumption in the ordinal response submodels of the mixed response model, covariate effects are allowed to vary across the grades. As a result, the final model allowed us to state that the conventional incubator specifically increases the numbers of top-grade embryos compared to the experimental incubator. By contrast, the proportional odds model only tells us that there is some overall advantage of the conventional incubator with respect to these responses. A disadvantage however is the fact that the grade-specific covariate effects are estimated with less precision than the overall effects in the proportional odds model. Jointly modelling the responses allows us to evaluate the latent correlation between morphology parameters using random effects. All three models give the same answer in this regard. We observed strong correlation between evenness and fragmentation degree, and modest correlation between cell number and fragmentation degree. Correlation between cell number and evenness appears to be negligible, or at best modest.

To investigate whether jointly modelling the responses offered any other advantages beyond the ability to estimate latent measures of association, we fitted the submodels in the mixed response model as three separate regression models (not shown). Both the model estimates and their corresponding precision were essentially unchanged. Results from simulation studies have arrived at different conclusions in relation to efficiency gains from joint modelling, with Lesaffre et al., (1991) and Gueorguieva and Agresti, (2001) reporting no substantive gains. By contrast, McCulloch, (2008) and Gueorguieva and Sancora, (2001) both reported impressive reductions in standard errors in submodels for binary and ordinal response variables when jointly modelled with a continuous response. Our results are in line with the former group of studies. One possibility is that the lack of efficiency gain stems from the inclusion of the same covariates in each submodel. In the case of seemingly unrelated regression (SUR, Zellner, 1962), where linear regression models are linked by correlated error terms, it has been shown that no efficiency gains are achieved for covariates shared across the submodels (Zellner and Huang 1962, Oliveira and Teixeira-Pinto 2015, Breiman, 1997). Our present models differ from the typical SUR models in that we accommodate correlation between the responses using random effects defined at the level of the treatment cycle (level 2) rather than at the level of the embryo (level 1). Intuitively however, the choice of covariates in the submodels may influence the efficiency of joint modelling in an analogous manner to the SUR case.

6.3 Summary of Chapter 6.

In this chapter, we have developed models for responses at the stimulation and embryo culture stages of IVF. In the latter case, we used joint modelling approaches to simultaneously analyse several embryo morphology outcome measures, thereby permitting the estimation of measures of association between them. In the next chapter, we extend these mixed response joint models to multiple stages of the IVF cycle.

6.4 Supplementary material for Chapter 6.

Predictor	Evenness Grade			
	1	<=2	<=3	<=4
Incubator (Embryoscope/Hunter)	1.19	1.13	1.11	1
Fertilization Method (ICSI/IVF)	1.44	1.10	1.06	1
Attempt No				
(Attempt 1/ Attempt 2)	0.81	0.92	0.98	1
(Attempt 1/ Attempt 3)	0.75	0.83	1.05	1
Age (Quantiles)				
(Q1/Q2)	1.78	1.20	1.02	1
(Q1/Q3)	1.04	1.20	1.01	1
(Q1/Q4)	1.02	1.27	1.04	1
Partner Age (Quantiles)				
(Q1/Q2)	0.93	0.98	1	1
(Q1/Q3)	0.95	1.13	1	1
(Q1/Q4)	1.37	1.17	1.01	1

S Table 16: Ratios of cumulative probabilities of embryos having a given day 2 evenness grade or lower according to predictor variables. The proportional odds assumption requires that values in each row are similar, excluding the final column.

Predictor	Fragmentation Degree			
	1	<=2	<=3	<=4
Incubator (Embryoscope/Hunter)	1.36	1.32	1.23	1
Fertilization Method	1.23	1.13	1.03	1
(ICSI/IVF)				
Attempt No				
(Attempt 1/ Attempt 2)	0.65	0.83	0.91	1
(Attempt 1/ Attempt 3)	0.76	0.78	0.83	1
Age (Quantiles)				
(Q1/Q2)	1.30	1.14	0.95	1
(Q1/Q3)	1.08	1.26	1.02	1
(Q1/Q4)	1.05	1.11	0.90	1
Partner Age (Quantiles)				
(Q1/Q2)	1.25	0.93	0.88	1
(Q1/Q3)	0.95	1.00	0.80	1
(Q1/Q4)	1.64	1.15	0.94	1

S Table 17: Ratios of cumulative probabilities of embryos having a given day 2 frag degree grade or lower according to predictor variables. The proportional odds assumption requires that values in each row are similar, excluding the final column.

6.5 References for Chapter 6.

Akaike H. Information theory and an extension of the maximum likelihood principle. Proc 2nd Int Symp Information Theory, Supp to Problems of Control and Information Theory 1972: 267-281.

Breiman L, Friedman JH. Predicting multivariate responses in multiple linear regression. J R Stat Soc Series B Stat Methodol. 1997 Jan 1;59(1):3-54.

Gelman A, Goegebeur Y, Tuerlinckx F, Van Mechelen I. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Appl Statist* 2000;1;49(2): 247-68.

Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat Sinica* 1996; 6: 733-760.

Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci* 1992: 457-472.

Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustering binary and continuous responses. *J Am Stat Assoc* 2001; 96 (455): 1102-12.

Gueorguieva RV, Sanacora G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat Med*. 2006;25(8):1307-22.

Harbottle S, Hughes C, Cutting R, Roberts SA, Brison DR, Association of Clinical Embryologists, British Fertility Society. Elective Single Embryo Transfer: an update to UK Best Practice Guidelines. *Hum Fertil* 2015; 18 (3): 165-183.

La Marca A, Sunkara SK. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Hum Reprod Update* 2014;20: 124-140.

Lesaffre E, Molenberghs G. Multivariate probit analysis: a neglected procedure in medical statistics. *Stat Med* 1991; 10: 1391-403.

McCulloch C. Joint modelling of mixed outcome types using latent variables. *Stat Methods Med Res* 2008; 17(1): 53-73.

O'Hara RB, Kotze DJ. Do not log-transform count data. *Methods Ecol Evol* 2010;1: 118-122.

Oliveira R, Teixeira-Pinto A. Analyzing Multiple Outcomes: Is it Really Worth the use of Multivariate Linear Regression? *J Biom Biostat* 2015;6.

Stan Development Team. Stan Modeling Language: User’s Guide and Reference Manual. 2017.

Su YS, Gelman A, Hill J, Yajima M. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *J Stat Softw* 2011;45: 1-31.

Zellner A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J Am Stat Assoc* 1962 Jun 1;57(298):348-68.

Zellner A, Huang DS. Further properties of efficient estimators for seemingly unrelated regression equations. International Economic Review. 1962 Sep 1;3(3):300-13.

Chapter 7. Methods for multistage models

In this section, we develop multistage methods for IVF data. We describe our thinking in developing these approaches more fully here than we do in Journal Article 5 (Chapter 10), which is a methodology paper arising from this work, although some overlap is unavoidable.

Our goal is to develop a framework for investigating the relationships between the responses at different treatment stages, as well as the impact of predictive variables on each of these. In the previous chapter, we tackled models relating to the ovarian stimulation and embryo culture stages of treatment. In this chapter, we consider the problem of how to link stage-specific submodels to produce a model for the full IVF cycle. The challenge is exacerbated by the mixed response variables we encounter (for example, binary, count and ordinal responses), and the fact that some of these are defined at the level of the embryo rather than the patient. We first consider a direct extension of our joint embryo models considered last chapter, and link submodels relating to responses at different stages using correlated latent variables. As limitations of this approach become apparent, we consider alternative approaches, where we allow the stage-specific response variables to enter into the downstream response submodels as covariates.

7.1 Motivating schematic

For the purpose of exposition, we begin with the schematic of the fresh IVF cycle displayed in Figure 27, where we present the IVF cycle as comprising three distinct stages; 1) stimulation of the ovaries, 2) fertilization and embryo culture, and finally 3) embryo transfer. We use the number of oocytes obtained as a measure of success of ovarian stimulation, three embryo quality scores (cell number, fragmentation and evenness) as indicators of success of the egg fertilisation and embryo culture stage, and live birth event as the standard of success for the embryo transfer stage. Singleton and twin births are both considered to be live birth events, and are not differentiated in the model. The outcomes number of eggs (count variable) and live birth event (binary) are defined at the level of the cycle, while the embryo gradings (one discrete that we treat as continuous and two ordinal measures) are defined for each individual embryo in the cycle.

This presentation is simplistic, and we anticipate that additional stages and responses will have to be included in order to answer real clinical research questions about the IVF cycle (see Chapters 10 and 11). For example, egg fertilization and embryo culture are really distinct stages, with the response variable 'number of embryos created' preceding measures of the quality of those embryos. By aggregating these stages, we are effectively ignoring the fact that the number of embryos produced is itself informative. Noting its practical inadequacy, we proceed with the three-stage representation for the purposes of exposition of the methods.

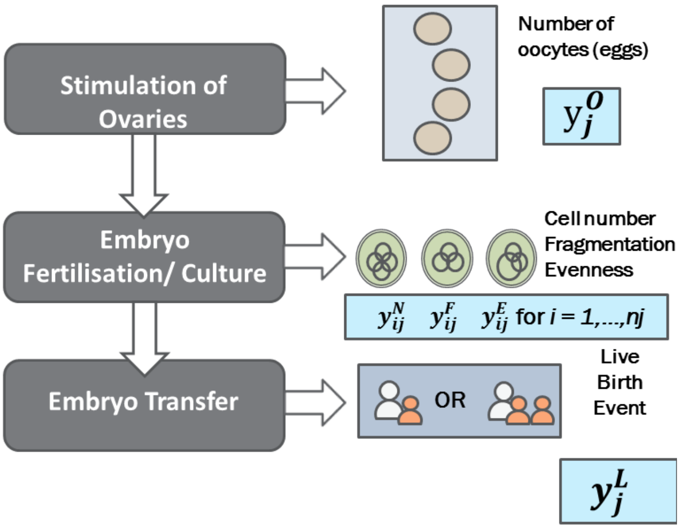


Figure 27: Schematic of the fresh IVF cycle for embryo i in cycle j. We jointly model outcomes at each of three stages.

7.2 Joint modelling using correlated latent variables

Our initial approach is a direct extension of the embryo models employed in the previous chapter. Whereas previously we defined regression submodels for the three embryo quality parameters, we now define additional submodels for the response variables preceding and following embryo culture. We estimate the relationship between outcomes by supposing that there is a multivariate Gaussian structure underlying the multistage data consisting of latent variables from each submodel (Goldstein, et al., 2009). The correlation matrix of this Gaussian distribution is then estimated along with the other model parameters.

7.2.1. Review of the submodels

We begin with a narrative description of the submodels, so that the reader may skip the subsequent mathematical presentation if desired. In light of the analyses conducted in the previous chapter, the cycle-level outcome 'number of oocytes' is modelled using an overdispersed Poisson regression model, with a cycle-level latent variable (alternatively, 'random effect') representing unmeasured covariables. We represent the embryo-level outcomes as in the mixed response models employed in the previous analyses. 'Cell number' is log transformed using a base of 2 as this can be interpreted as the number of doublings. Since embryos are nested within cycles, log2(cell number) is then modelled using a standard two-level linear mixed model, including a random intercept term to capture between-cycle heterogeneity. The ordinal embryo gradings 'evenness' and 'fragmentation degree' are modelled using cumulative logit models, again with random intercept terms representing between-cycle heterogeneity. These random terms can be viewed as latent variables, and it is these that are used to model the relationship between outcome measures. The cycle-level outcome 'live birth event' is modelled using a latent variable representation of a probit model, where positive (negative) values of the latent variable correspond to a success (failure). The cycle-level latent variable is modelled using linear regression. The error term from this latent model is used to estimate the correlation with the responses at earlier stages of the cycle.

7.2.2. Mathematical representation of the submodels

Stimulation phase

In the current presentation, we are ignoring any clustering arising from repeated treatment cycles as relatively few patients in the dataset (described below) have these. For cycle j, we assume the number of oocytes obtained y_j^O follows a Poisson distribution and model the log of the rate parameter λ_j^o in the usual way:

$$\log(\lambda_j^o) = \boldsymbol{S}_j \boldsymbol{\beta}_o + z_j^o$$

where $\boldsymbol{S}_j$ is a row-vector of cycle-level covariates for cycle *j*, $\boldsymbol{\beta}_o$ is a corresponding vector of regression parameters and z_j^o is a cycle-specific latent variable that models overdispersion in the oocyte yield. This latent term is used to capture the relationship between the stimulation response and outcomes at later stages, as described below.

Embryo fertilization and culture

Our embryo submodels display the same form as those employed in the previous chapter, and we repeat the specification here because it facilitates the exposition and is quite brief in any case. For embryo i (where i =1,…,n$_j$) nested in cycle j we model y_{ij}^N, log$_2$ (cell number), using a two-level linear regression model:

$$y_{ij}^N \ = \boldsymbol{X}_{ij}\boldsymbol{\beta}_N + z_j^N \quad + e_{ij}^N$$

$$e_{ij}^N \sim N(0, \sigma_N^2)$$

We use cumulative logit models for the ordinal outcomes. For k = 1,2,3:

$$\text{logit}\left(\gamma_{kij}^E\right) = \ \alpha_{kE} - \ \boldsymbol{X}_{ij}\boldsymbol{\beta}_{kE} - z_j^E$$

$$\text{logit}(\gamma_{kij}^F) = \ \alpha_{kF} - \boldsymbol{X}_{ij}\boldsymbol{\beta}_{kF} - z_j^F$$

Where $\boldsymbol{X}_{ij}$ is a row-vector of covariates, $\boldsymbol{\beta}_N$, $\boldsymbol{\beta}_{kE}$ and $\boldsymbol{\beta}_{kF}$ are each vectors of regression coefficients, and z_j^N , z_j^E and z_j^F are cycle-level random effects (latent variables). γ_{kij}^E and γ_{kij}^F are cumulative probabilities of embryo *i* in cycle *j* having a grade of *k* or lower for evenness and fragmentation degree respectively and α_{kE} and a_{KF} $(k = 1{,}2{,}3)$ are threshold parameters, corresponding to the log-odds of the embryo having this or a lower grade.

Live birth event

We use a latent variable representation of a probit regression model for the clinical outcome of the cycle, live birth event. Let $y_j^L = 1$ or 0 if cycle does or does not result in a live birth, respectively. We define y_j^{L*} as a latent continuous variable underlying the binary y_j^L, such that

$$y_j^L = \begin{cases} 1 \ if y_j^{L*} \ \geq 0 \\ 0 \ if \ y_j^{L*} \ < 0 \end{cases}$$

A linear regression model for the latent y_j^{L*} is then used to estimate covariate effects:

$$y_j^{L*} = \boldsymbol{C}_j\boldsymbol{\beta}_* + z_j^*$$

$$z_j^* \sim N(0, 1)$$

where $\boldsymbol{C}_j$is a row-vector of cycle-level covariables and $\boldsymbol{\beta}_*$ is a r x 1 vector of regression coefficients. Fixing the variance of z_j^* to be 1 is mathematically equivalent to specifying a probit model for the probability that a single transfer cycle culminates in a live birth event. We choose a probit model over the more familiar logistic regression approach in order to allow the correlation between LBE and the embryo parameters to be estimated. This is achieved by specifying a multivariate normal distribution for the cycle-level random terms appearing in the first four submodels and the latent error term in the fifth:

$$\begin{bmatrix} z_j^O \\ z_j^N \\ z_j^E \\ z_j^F \\ z_j^* \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \theta_O^2 & \eta_1\theta_O\theta_N & \eta_2\theta_O\theta_E & \eta_3\theta_O\theta_F & \eta_4\theta_O \\ \eta_1\theta_N\theta_O & \theta_N^2 & \eta_5\theta_N\theta_E & \eta_6\theta_N\theta_F & \eta_7\theta_N \\ \eta_2\theta_E\theta_O & \eta_5\theta_E\theta_N & \theta_E^2 & \eta_8\theta_E\theta_F & \eta_9\theta_E \\ \eta_3\theta_F\theta_O & \eta_6\theta_F\theta_N & \eta_8\theta_F\theta_E & \theta_F^2 & \eta_{10}\theta_F \\ \eta_4\theta_O & \eta_7\theta_N & \eta_9\theta_E & \eta_{10}\theta_F & 1 \end{bmatrix}\right)$$

The elements η_1, …, η_{10} of the vector $\boldsymbol{\eta}$ represent latent correlation coefficients, and act as a measure of association between response variables.

7.2.3. Description of the dataset

The data used in this example are all fresh IVF cycles started between June 2013 to June 2014, at St Mary's Department of Reproductive Medicine, Manchester, where patients used their own eggs. In total, this represents 1091 treatment cycles. Embryo outcome data were only available for a subset of the cohort (634 cycles). This isn't ideal but we ignore it for present purposes. A small proportion of these cycles are repeated treatments on the same patients, but this is also ignored in the present example and we use 'cycle' and 'patient' interchangeably in the following.

The rate of drop-out at each stage of the cycle was rather low. Eighty-seven (8%) cycles did not proceed beyond the stimulation and egg collection stage, 58 (5%) did not proceed beyond the fertilization and embryo culture stage, with the remaining 946 (87%) of cycles resulting in transfer. There were 318 live birth events, giving a LBE per cycle started rate of 29% and a LBE per transfer procedure rate of 34%.

7.2.4. **Model fitting**

We fit the joint model as a Bayesian hierarchical model. The model is hierarchical due to the underlying multivariate Normal distribution which can be thought of as a prior distribution on the latent variables, with the parameters of this prior estimated from the data. This multivariate Normal prior then has a hyperprior distribution, which we specify using a LKJ(1) distribution. This is uniform over all of the correlation matrices of appropriate dimension (Stan Development Team, 2017). This may be less efficient than placing a prior on the Cholesky factorisation of the correlation matrix (Stan Development Team, 2017), but we do not pursue this yet. We place noninformative priors on most parameters. Exceptions are the priors for regression coefficients for the LBE submodel. The latent variables in this submodel are scaled to have a variance of 1, and it is therefore rather unlikely that the coefficient values would be much greater than this. We therefore use Normal(0,2) priors for these parameters, which amount to being very weakly informative. The primary motivation here is to speed up the algorithm used to fit the model. We run two chains for 3000 iterations, discarding the first half as a warm up period. This leaves 3000 draws from the posterior distribution for the purpose of inference, and we conduct standard convergence diagnostics before proceeding (not shown). In a genuine application, we would run additional chains and conduct extensive model checks via the posterior predictive distribution to assess the fit.

7.2.5. **Covariates in the model**

We include three baseline covariables in this example (age, partner age and attempt number). We include these in all of the submodels, although there is no requirement for each submodel to include the same predictor variables in general. In this simplistic example, we include these variables without considering the most appropriate form that they should take. For example, the relationship between age and stimulation response is likely to be nonlinear. We do not consider these points here, although appropriate representations could be investigated via the usual exploratory data analysis.

7.2.6. Dropout from the cycle

We have described dropout resulting from treatment failure as a complicating factor in the analysis of multistage IVF data. A patient may have an inadequate ovarian stimulation response, for example, preventing progression to subsequent treatment stages. We fit our submodels conditional on 'success' (which here we use to mean 'not failure', rather than a good outcome) at the previous stage. For example, a patient who failed (or otherwise dropped out) during the stimulation phase would not provide direct information on either the embryo outcomes or transfer outcome, as they lack outcome data for these stages. They would however provide some information via their available data – namely through their covariate data and their response to ovarian stimulation. Latent variables corresponding to the unrealised treatment stages are drawn for patients who drop out although, as is usual in multilevel modelling, they are shrunken towards the mean for patients with similar values for their observed data (Gelman, et al., 2012). One could attempt to interpret these uncoupled latent variables as propensities to produce embryos of a given size, evenness etc, but this might require too far a stretch of the imagination. For now, we note that the latent variables for unrealised stages are drawn under the assumption that, whatever they represent, they assume that the relationships between stages are the same for those who do and do not drop out. We discuss the matter of how to handle dropout in the multistage context in more detail in Chapter 8.

7.2.7. Results of fitting the correlated latent variable model

We do not go into detail about the estimated values of the regression coefficients from the various submodels here, as they have their usual interpretation. We focus instead on the estimated correlation matrix and the interpretation of it. The estimated correlation matrix is shown in Table 11. It suggests that patients obtaining more eggs after stimulation have larger embryos (greater cell number). It also suggests that patients with more eggs and larger embryos are more likely to have a successful embryo transfer.

Conventional wisdom suggests that an increased number of oocytes increases the chance of overall success due to the greater pool of embryos from which to select. However, these results could generate the hypothesis that patients with increased oocyte yield actually have superior embryos, and this may contribute to transfer success. Fragmentation Degree is strongly associated with evenness and weakly (or at most moderately) associated with cell number.

7.2.8. How to interpret latent correlation coefficients?

Superficially, it appears that an estimated latent correlation coefficient can be given a reasonably straightforward interpretation, provided that the two submodels to which it refers share the same covariables. If so, the correlation coefficient can be interpreted as a measure of association adjusted for those covariables. This can be understood by way of analogy with multiple linear regression. Suppose that an outcome variable *y* is regressed on several covariates (*x1, x2, x3,* say). The regression coefficient corresponding to *x1* can be recovered by carrying out the following three steps. First, regress *y* on both *x2* and *x3*

	Number of oocytes	Log_2(Cell Number)	Evenness	Fragmentation Degree	Live Birth Event
Number of oocytes	1	0.26 (0.07 to 0.43)	0.13 (-0.03 to 0.29)	0.09 (-0.05 to 0.24)	0.31 (0.16 to 0.46)
Log_2(Cell Number)	0.26 (0.07 to 0.43)	1	0.02 (-0.11 to 0.15)	0.18 (0.06 to 0.30)	0.30 (0.13 to 0.47)
Evenness	0.13 (-0.03 to 0.29)	0.02 (-0.11 to 0.15)	1	0.86 (0.81 to 0.91)	0.05 (-0.12 to 0.21)
Fragmentation Degree	0.09 (-0.05 to 0.24)	0.18 (0.06 to 0.30)	0.86 (0.81 to 0.91)	1	0.03 (-0.13 to 0.18)
Live Birth Event	0.31 (0.16 to 0.46)	0.30 (0.13 to 0.47)	0.05 (-0.12 to 0.21)	0.03 (-0.13 to 0.18)	1

Table 11: Estimated correlation matrix (posterior means and 95% CIs) from the correlated latent variable model.

(including both as covariates in a multiple regression). Second, regress *x1* on both *x2* and *x3* (again, including both as covariates in a multiple regression). Finally, regress the residuals from the first step on the residuals from the second. The slope of this regression line will duplicate the regression coefficient for *x*1 in the regression of *y* on *x1*, *x2* and *x3*, and can be interpreted as an estimate of the association between y and x1 after adjusting for x2 and x3.

The joint model of two outcomes (or more accurately for generalized linear submodels, of their transformed means) can be similarly construed if the covariates in the two models are identical. For example, consider a joint model of $\log_2$(cell number) and number of oocytes, containing two baseline covariates. The former is modelled using linear regression. The latter is modelled using Poisson regression with the log rate parameter expressed in terms of the linear predictor and a random (cycle-varying) term representing unmeasured covariates. The residuals in the first submodel and the random term in the second are analogues of the residuals from the regressions of *y* and of *x1* on the covariates *x2* and *x3* in the multiple linear regression example above. In the multiple linear regression example, we regressed one set of residuals on the other. In the joint model, we instead estimate the correlation. In both cases, we assess the relationship between the unexplained variation from the first submodel and the unexplained variation from the second, and can consider this estimate of association to be adjusted for the other covariates.

Although such an interpretation is possible, (and possibly useful), it falls short of a measure of association on the scale(s) of the response variables in the model. We return to this point later.

7.2.9. Choosing covariates for causal inference and for prediction

Thinking about the correlation coefficients from the joint model in this way suggests some principles relating to covariate selection. If we hope to assign a causal interpretation to the estimates of association, then it would not be appropriate to include covariates that lie on a causal pathway intermediate to the conjoined outcome measures under consideration. This scenario would essentially be precluded by the requirement that both submodels include the same covariates, since it would lead to the inclusion of a covariate that occurred subsequent to the outcome variable in one of the submodels. Clearly, this would be nonsensical. If there is interest in interpreting the correlation coefficients as causal, then the model covariates should be the same in all submodels and must precede each of the outcomes in the model. Yet this might not be enough to permit a causal interpretation. Even when the response variables are temporally ordered, and have been appropriately adjusted by covariates, the submodels in this framework cannot be considered to be adjusted for the other response variables in the joint model. For example, the relationship between cell number and live birth in Table 11 could in principle be partially or wholly attributable to effects of number of oocytes on both of these variables. If so, we would have confounding of the relationship by the number of oocytes obtained (here, we are setting aside prior plausibility and noting that this interpretation would be consistent with the model). Furthermore, a correlation coefficient, even if adjusted, can never be given a meaningful causal interpretation, since its magnitude depends on the variation in the sample (Greenland, et al., 1991). As such, it cannot be taken as a measure of a stable law of nature. If interest lies in investigating how a response at one stage affects what happens next, it isn't clear that this approach provides the means by which to do it.

On the other hand, the intention may be to use the model for the purposes of prediction. The model may be particularly useful specifically for inherently multivariate prediction problems. For example, a joint model could be used to predict the chance that a patient with certain characteristics treated under a certain stimulation protocol might have a safe stimulation response (a yield of eggs falling in an acceptable range) and then go on to a successful embryo transfer. This might be useful when trying to design optimal ovarian stimulation strategies, where there is a need to balance effectiveness and safety. This is likely to offer advantages over approaches based on composite outcomes, which waste information and do not allow for differential effects of covariates on safety and effectiveness. For pure prediction problems, we are less concerned about violating the conditions for valid causal inference, and may choose which covariables to include on more pragmatic grounds (such as the availability or feasibility of gathering the measure in clinical practice, or the fact that it improves predictive performance of the model, see eg: Steyerberg, 2008).

7.2.10. Assumptions of the model

The most substantial assumption we make in this approach is the assumption that, conditional on the covariates and on the latent variables that we use to accommodate correlation between the outcome measures, the outcomes being modelled are independent (Gueorguieva, 2001, McCulloch, 2008). That is, if we have two response variables *y1* and *y2*, with a covariate *x* and response-specific latent variables *z1* and *z2*, we assume that

$$p(y1, y2|x, z1, z2) = p(y1|x, z1, z2)p(y2|x, z1, z2)$$

To understand this point we can make an analogy with longitudinal data analysis, where we usually assume that repeated observations on an individual are independent conditional on the random effects in a mixed model. In the LDA setting, we can think of the random effects as representing unmeasured time-invariant covariates that account for the unexplained heterogeneity between participants. A consequence of this for LDA is that, unless we introduce greater complexity through the observation-level residuals, we assume that the correlation between repeated measures is constant. By contrast our approach, where correlation is built in by placing a multivariate Normal distribution on the random effects or residuals from each submodel, allows for considerable flexibility in the covariance structure underlying the multivariate response; no assumption of constant correlation between all of the response variables is required. One inconvenient consequence of the conditional independence assumption in our model however relates to embryo gradings. In particular, it entails that the correlation between two measures of embryo grade for any particular embryo would be the same if we were to take the same two measures individually from two embryos belonging to the same patient (Gueorguieva, 2001). This is unrealistic, and it would be desirable to relax it.

The coherence of the conditional independence assumption for two response variables can be assessed by including one as a covariate in the submodel for the other (Gueorguieva, 2001). A coefficient of zero would correspond to conditional independence. In our case, we expect the outcomes at each stage to play an important role in determining ‘downstream’ outcomes. It is probably unrealistic to assume that this dependence can be fully accounted for through the multivariate Normal distribution we specify for the latent variables. The degree of model misspecification is unknown in practice.

At this point we also note a second assumption of this approach. In the absence of an explicit model for drop out, we essentially assume that unobserved outcomes can be explained as missing at random (MAR, Rubin, 1976). This means that we assume that the missing responses are ignorable given the observed data included in the model. There may be some plausibility to this assumption, since the response at each stage largely determines whether or not a patient continues the cycle, and response variables representing key milestones are included in the joint model. The dependency between response variables that throws the conditional independence assumption into doubt therefore also increases the plausibility of missingness at random. We return to the matter of modelling the dropout process in Chapter 8.

7.2.11. Improving the correlated latent variable approach

The latent variable approach outlined here has limitations. Given the anticipated strong dependence between response variables, the conditional independence assumption may be untenable. Moreover, if our goal is to understand efficacy and mechanism of IVF interventions, it would be useful to be able to estimate the effects of upstream outcomes on downstream responses. While the correlation coefficients from our approach may be adjusted for confounding variables, it would be preferable to obtain measures of association on the same scales as the model response variables. In the next section, we attempt to overcome these limitations by considering models where response variables are included as covariates in the submodels for downstream outcomes.

7.3 Including procedural responses as covariates in submodels for downstream responses

The approach based on estimating the correlation between cycle-level latent variables has limitations (section 7.2). Although it is possible to obtain adjusted estimates of association between responses at different stages of the cycle using this approach, the interpretation of these estimates is somewhat obscure. An approach that allows valid and directly interpretable estimates of relationships between response variables, the endogenous response model, is developed here.

7.3.1. Motivating example: a two-stage IVF model

In order to introduce the endogenous response modelling approach, we begin by considering a simple two-stage model of the IVF cycle, featuring only the stimulation and transfer stages (so that we do not include the fertilisation and culture stage/stages at present). This simplifies the situation compared to the schematic portrayed in 7.1 and Figure 27 since it leaves us with responses measured at a single level (the level of the cycle). Again, we ignore repeated cycles undertaken by the same patients here and treat the cycle as the unit of analysis. As with the previous example, we use the number of oocytes retrieved as the outcome measure of stimulation response and live birth event as the outcome measure corresponding to embryo transfer. We again consider only a small number of baseline covariables (age, partner age and attempt number).

We are interested in the effects of covariates on stimulation response and on transfer success. Moreover, we are interested in the relationship between the oocyte yield and the likelihood of a successful embryo transfer.

7.3.2. Correlated latent variable two-stage model

We first define a correlated latent variable model for the two-stage scenario. As for the three-stage model presented in 7.2, a correlated latent variable model for the two-stage schematic might contain an overdispersed Poisson submodel for the number of oocytes obtained from ovarian stimulation, and a latent probit model submodel for live birth event. The correlation between the two response variables could then be accommodated by specifying a bivariate Normal distribution for the latent variables:

$$\begin{pmatrix} z_j^O \\ z_j^* \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \theta_O^2 & \eta\theta_O \\ \eta\theta_O & 1 \end{pmatrix}\right)$$

This allows us to estimate the correlation between the stimulation response and outcome of transfer, η. We include the same covariates in both submodels, so that η can be interpreted as a measure of association adjusted for these variables.

7.3.3. Introducing outcome-covariates to the joint model

The endogenous response model can be thought of as an extension to the latent variable approach. In the endogenous response model, we allow procedural response variables to enter into the downstream response submodels as covariates. Wherever a response variable features as a covariate for another response, we refer to it as an outcome-covariate. In the two-stage scenario, we define an endogenous response model by including the stimulation response variable ‘number of oocytes’ as an additional covariate in the live birth event submodel. We still specify a multivariate Normal distribution for the latent terms in the two submodels, so that they are correlated.

This formulation is similar to the approach described by Terza (1998), who presented a joint model for estimating the effect of car ownership on household trip-taking. In that example, a straightforward overdispersed Poisson regression of number of household trips on vehicle ownership (and several other covariates) would not give a valid estimate of the treatment effect, due to the fact that vehicle ownership is likely to be correlated with other unmeasured sources of variation represented by the latent overdispersion term. Terza gives the example of the household’s attitude towards public transport as an unobserved variable which is likely to be correlated with both the number of trips taken and vehicle ownership. In the event that a covariate is correlated with the unobserved part of the model, we say that the covariate is *endogenous*. Terza jointly models the binary response ‘car ownership’ by representing this variable using a latent probit submodel, and allowing it to be correlated with the overdispersion term in the household trips submodel. The variable ‘car ownership’ also appears as an outcome-covariate in the latter. Assuming that the model specification is correct, this approach produces a valid estimate of the endogenous treatment effect because it incorporates the correlation resulting from unmeasured confounders.

The main difference between the example of Terza and the present case is that our interest is in estimating the effect of a count variable upon a binary response, rather than the other way around. In our IVF example it is highly plausible that the variable ‘number of oocytes’, if included as a covariate in the submodel for live birth event, would be endogenous. This is because unobserved variables relating to the health of the patient (for example) are expected to be related to both stimulation response and transfer success. Standard regression approaches for binary dependent variables would therefore not allow for the valid estimation of the effect of number of oocytes upon transfer success, because confounding due to unobserved factors is expected. By contrast, the endogenous response joint modelling approach allows for the endogeneity of ‘number of oocytes’ to be explicitly accommodated in the model.

7.3.4. Path diagrams for causal models

To demonstrate and compare the key features of these approaches, it will be useful to introduce path diagrams. We follow the conventions adopted by Skronal and Rabe-Hesketh (2004). Variables are represented by nodes in the diagram. We use rectangles to represent observed variables and circles to represent latent variables. Triangles represent constant terms (such as the ‘1’ representing an intercept term in Figure 28). Arrows represent directed relations (which we intend to interpret as causal, given some assumptions). Bidirectional arrows connecting nodes represent non-directed relations (eg: correlations). In some presentations variances are denoted by bidirectional arrows originating and terminating at the same node. We supress variances here in an attempt to prevent overcrowding as we introduce more complex models. Note that we use directional arrows regardless of the distribution of the included variables (see for example the arrows describing the nonlinear relationships between covariates and the responses number of oocytes (count) and live birth event (binary) in Figure 28). Arrows are labelled with the parameter representing the relation in the model. The enclosed box and the text ‘Cycle j’ indicate that the variables it contains vary between cycles, removing the need to include the ‘j’ subscript (Skrondal and Rabe-Hesketh, 2004).

Figure 28 shows a path diagram for the correlated latent variable model representing the situation described above, while Figure 29 shows the endogenous response approach. For convenience, we use the same parameter names for the corresponding relations in the two models, although these will not coincide. The sole difference between the two figures is an arrow originating from number of eggs and ending at live birth event, corresponding to the inclusion of number of eggs as an outcome-covariate in the live birth event submodel.

7.3.5. A comparison of the correlated latent variable and the endogenous response methods.

A comparison of

Figure 28 and Figure 29 highlight some key implications of treating number of eggs as an endogenous response. Firstly, it allows us to obtain an interpretable (or at least standard) estimate of the effect of number of eggs upon live birth event, adjusted for the other covariates in the submodel (β_{L6} in Figure 29). The coefficient of a probit regression corresponds to the change in Z score (where Z is a standard Normal variate, not to be confused with the latent variables denoted by $\boldsymbol{z}_j$ throughout the thesis). As for the coefficients derived from logistic regression models, they do not represent the 'marginal' effect of changing the covariate on the response variable and must be translated by setting the other covariates in the model to fixed values. Despite the slight effort required, the ability to interpret the coefficients in this manner puts the endogenous response approach at an advantage over the correlated latent variable approach, where we rely on the correlation parameter η to represent the relation between the response variables.

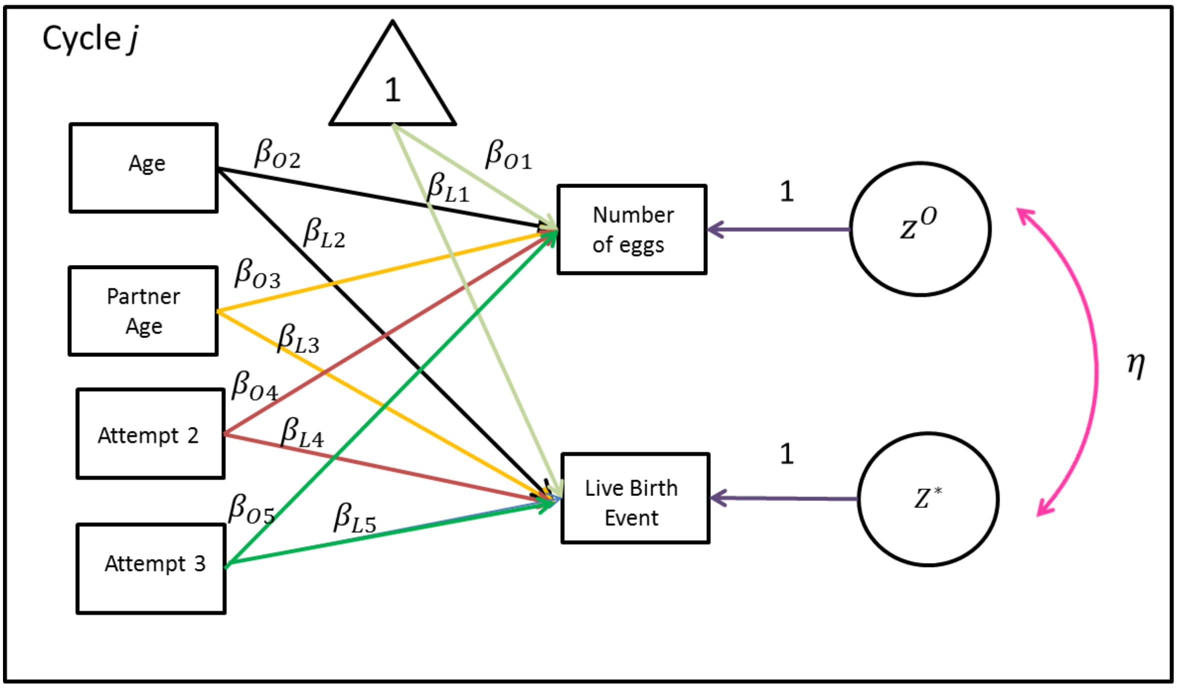


Figure 28: Path diagram showing a joint model of live birth event following embryo transfer and number of eggs obtained after ovarian stimulation (correlated latent variables approach). Directional arrows are coloured according to their origin for clarity.

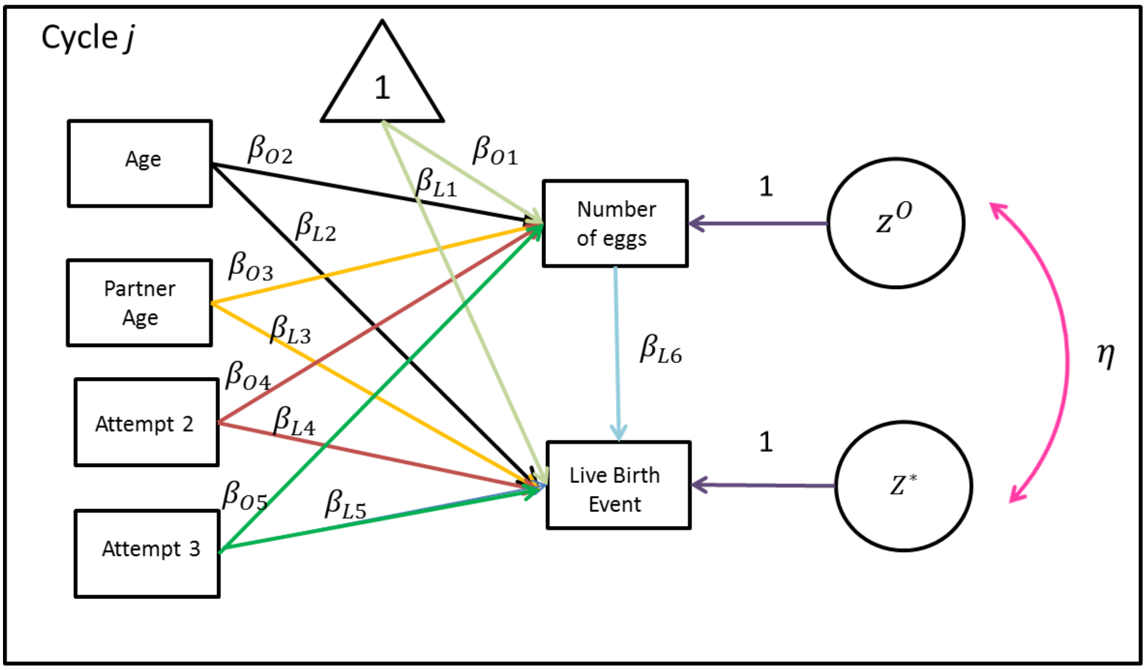


Figure 29: Path diagram showing a joint model of live birth event following embryo transfer and number of eggs obtained after ovarian stimulation (endogenous response approach). Directional arrows are coloured according to their origin for clarity.

7.3.6. Covariates in the endogenous response model

One concern in relation to including a response variable as a covariate in a regression model for a downstream response variable is that we might invalidate our estimates corresponding to upstream variables. For example, the number of oocytes obtained lies on the causal pathway from the baseline covariable age to the outcome live birth event[3]. Accordingly, adjusting for number of oocytes in a standard (univariate) probit regression model would invalidate the estimate of age. However, this is not the case in the joint model presented here, because the baseline covariables appear in both submodels. Figure 29 shows the implications of this formulation; in the endogenous response model the baseline covariables are permitted to act on live birth event indirectly through their effect on number of oocytes, and also ‘directly’ via some other pathway or pathways. Age, for example, could influence the probability of a transfer leading to a live birth by influencing the uterine environment, which is distinct from effects on eggs and embryos.

We could include certain covariables in one of the submodels, but not the other. For example, we could include age in the stimulation submodel but not in the transfer

[3] We discuss the matter of whether or not it is meaningful to speak of effects of non-manipulable variables later in the chapter.

submodel. This would imply that β_{L2} in Figure 29 is equal to zero, which means that there is no effect of age on transfer, other than indirectly by way of an effect on the number of oocytes obtained from ovarian stimulation. An alternative description would be that there is no *direct effect* of age on live birth event in the model (Pearl, 2001). However, given the simplistic nature of the model under consideration, this would include any effects of age on live birth event which were not mediated through the number of oocytes obtained. Given our interest in mechanistic models, the possibility of apportioning effects of predictors to different pathways is an attractive one.

7.4 Application to a three-stage characterization of the IVF cycle

In order to compare the correlated latent variable approach and the endogenous response approach in practice, as well as to investigate the feasibility of conducting the latter, we return to the three-stage representation of the IVF cycle presented in 7.1 (Figure 27). This includes an intermediate stage between ovarian stimulation and embryo transfer, corresponding to embryo fertilization and culture, and incorporating the three embryo response variables described in section 6.2 ($\log_2$(cell number), cell evenness and fragmentation degree). The representations of the five response variables in the model remain unchanged compared to section 7.2.2. Results of a correlated latent variable analysis were presented in Table 11. Here, we add the cycle-level response 'number of oocytes' as a covariate in the embryo submodels, with the implication that this variable is treated as endogenous. We do not include the embryo outcomes as covariates in the transfer submodel in this example, since these are only available for a subset of the cycles in the present dataset. When we do introduce embryo responses as outcome-covariates in the transfer submodel later on, we will be faced with the question of how exactly this should be done; since the embryo parameters are defined at a lower level of a multi-level data structure than the outcome of transfer. This point doesn't concern us here, and we note that the embryo responses are treated as being exogenous (that is, not endogenous) for now. A path diagram for the model is shown in Figure 30.

7.4.1. Fitting the endogenous response model

We again use rstan to fit the model (Stan Development Team). We place essentially flat $N(0,1000^2)$ priors on the regression coefficients, with the exception of those in the latent probit live birth event submodel, for which we use weakly informative $N(0,2^2)$ priors. This is as described for previous examples. Here, we use a Cholesky parametrisation of the correlation matrix to improve convergence (Stan Development Team, 2017). We again use an LKJ Correlation (1) prior, which is uniform over all possible correlation matrices, and a weakly informative Cauchy(0,2.5) prior for each of the scale parameters in the covariance matrix. We run 3 chains for 2000 iterations each, discarding the first half of these as burn in.

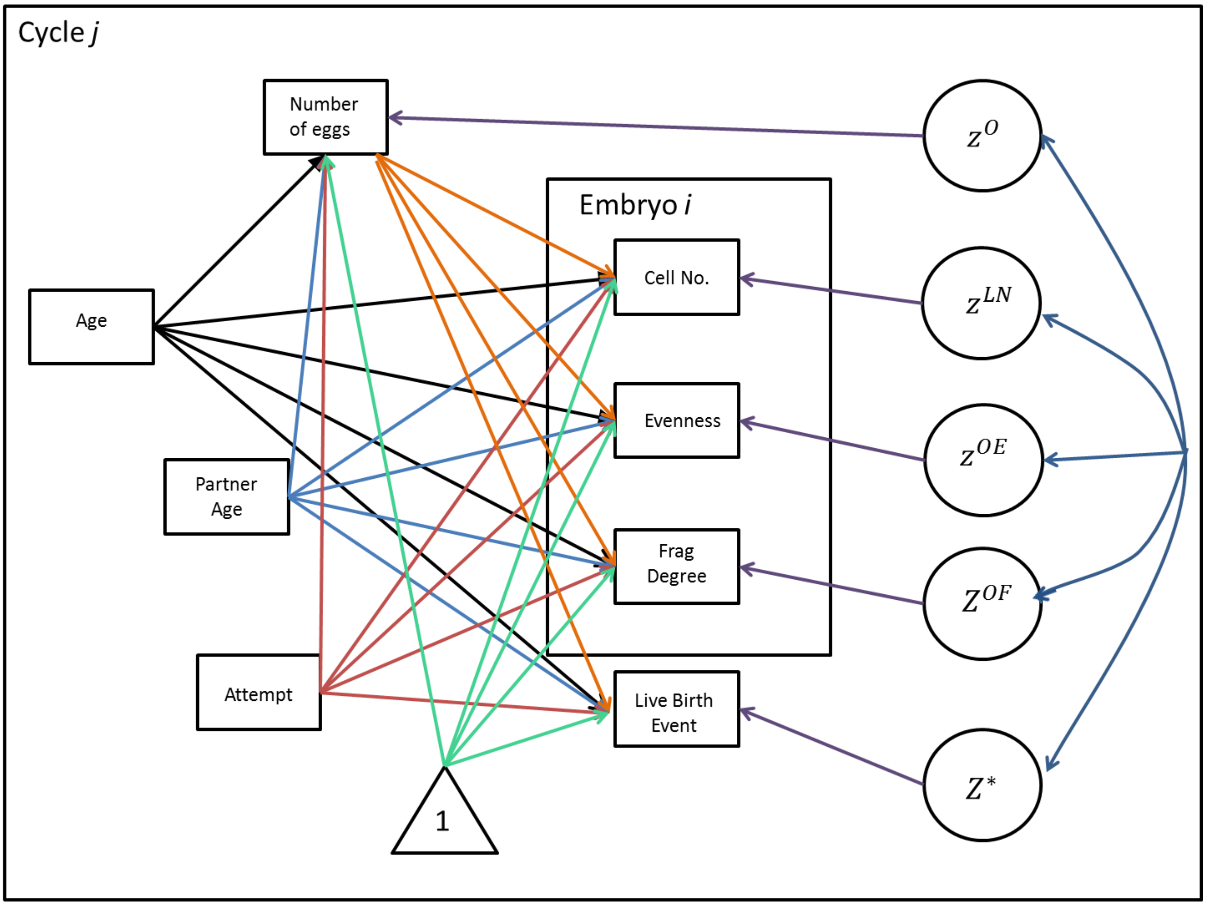


Figure 30: Path diagram showing a joint model of live birth event following embryo transfer, embryo quality variables and number of eggs obtained after ovarian stimulation. Number of eggs is included as a cycle-level covariate in each of the submodels for the downstream response variables. Directional arrows are coloured according to their origin for clarity. Paths have not been labelled with parameter names to avoid further cluttering an already busy display.

7.4.2. Output and interpretation of the two models

Journal Article 5 (Chapter 10) discusses differences in parameter interpretation between correlated latent variable and endogenous response approaches. Rather than duplicate that discussion, we focus here on the technical matter of how to obtain valid draws from the joint posterior distribution.

7.4.3. Convergence in endogenous response models

Several authors have noted that identifying and fitting endogenous variable models can be challenging in practice (Diggle, et al., 2007, McConnell, et al., 2008, Steele and Washbrook, 2013, Xie, 2000). Weak identification of the model parameters typically manifests as poor convergence of the sampling algorithm. We assess convergence of the model using standard techniques. Although the Gelman-Rubin convergence statistics are acceptable for all of the parameters in the model (all less than 1.1), our effective sample sizes are low (Gelman and Rubin, 1992). For example, we end up with effective sample sizes between 20 and 70 (out of a possible 3000) for the elements of the correlation matrix, and similarly low values for several of the regression coefficients in the embryo submodels. This suggests that the chains are slow to move around the posterior distribution. We can visualise this by looking at autocorrelation plots for the model parameters. An example is given in Figure 31, which shows the autocorrelation for the elements of the correlation matrix. The correlation remains high after 25 lags. In fact, the autocorrelation does not drop to acceptable levels within 150 lags for some of the parameters (not shown). This represents a practical obstacle, as it means that the model must be run for a long time on standard hardware before we obtain sufficient draws from the posterior distribution. It may be possible to improve convergence speed by changing the parametrisation of the model.

7.4.4. Improving mixing in the responses as covariates model: three possible approaches

Improving mixing through reparameterisation

One way to improve convergence speed might be to reparameterise the model (Rstan manual). To assess this, we adopt a shared parameter approach to joint modelling mixed outcomes similar to that discussed by Gueorguieva (2001) and McCulloch (2008). This involves including a common latent variable in the submodels for each response variable, and scaling this by a constant (to be estimated from the data) in all but one of these. Dunson and colleagues presented Bayesian formulations of shared parameter models (Dunson, 2000, Dunson, et al., 2003, Dunson and Herring, 2005) including one example where outcomes (birthweight and litter size) were measured at different levels (individual mice pups and the litter) of a multilevel data hierarchy (Dunson, et al., 2003). We adopt a similar approach here, including a common cycle-level latent variable δ in each of the response submodels. The latent variable is scaled in each submodel by a unique factor loading (λ_d, d = 1,2,3,4,5) which is estimated from the data. A path diagram representing this approach is displayed in

Figure 32

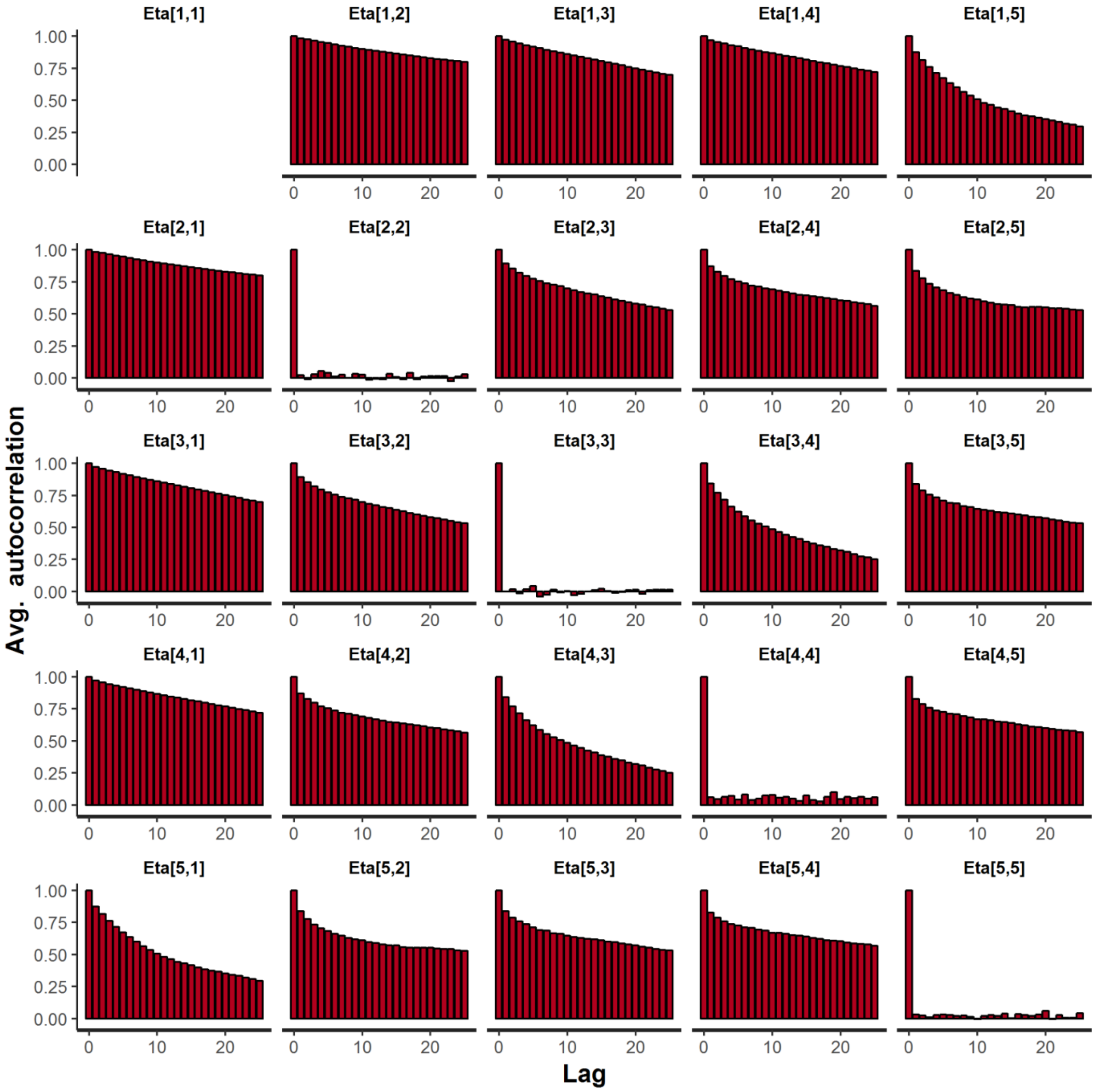


Figure 31: Autocorrelation plot showing the correlation between parameter draws at lag x, for each element of the coefficient matrix in the joint model of ovarian stimulation, embryo culture and embryo transfer which includes the variable 'number of oocytes', representing response to ovarian stimulation, as a covariate in the submodels for the downstream stages of treatment.

.

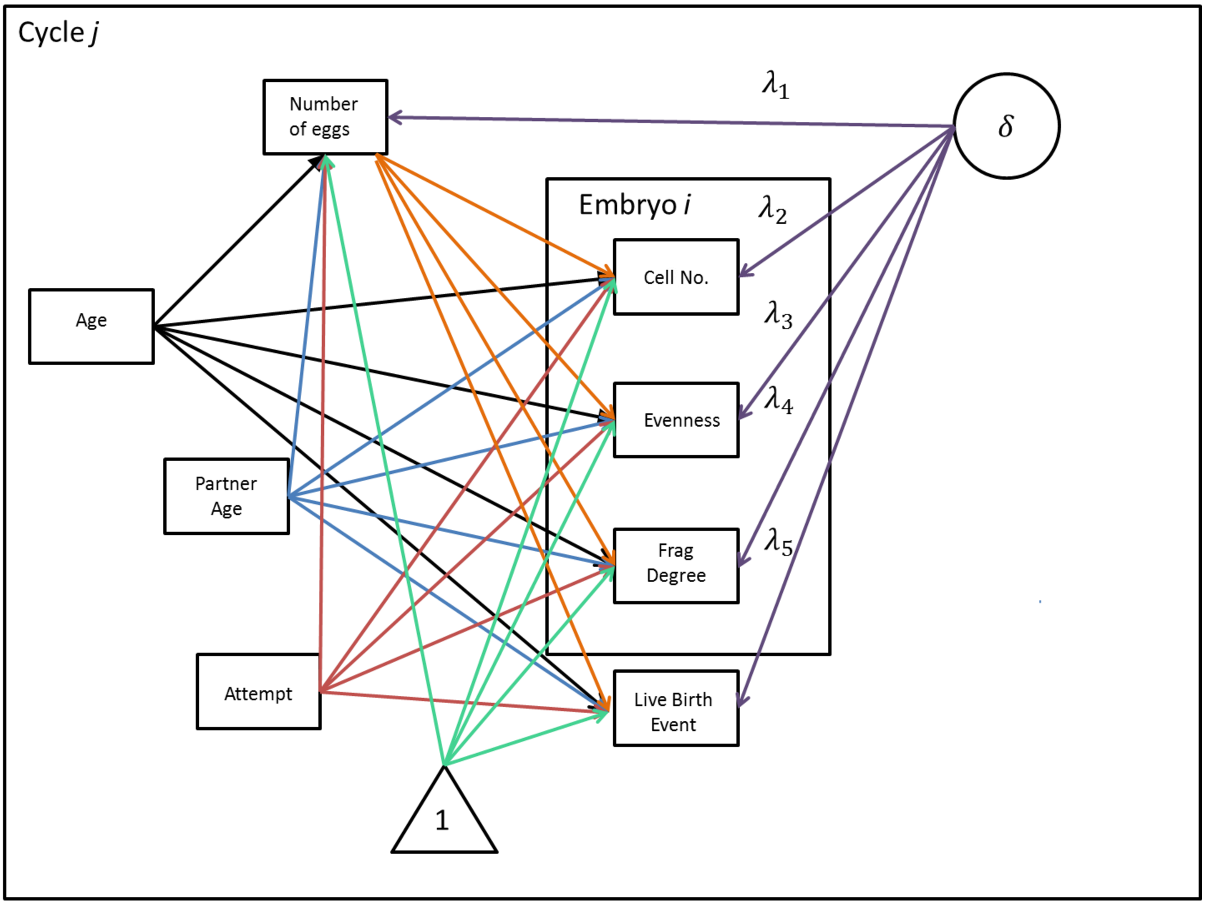


Figure 32: Path diagram showing a shared latent variable representation of a joint model of live birth event following embryo transfer, embryo quality variables and number of eggs obtained after ovarian stimulation. Number of eggs is included as a cycle-level covariate in each of the submodels for the downstream response variables. Directional arrows are coloured according to their origin for clarity. Other than those originating from the shared latent variable, paths have not been labelled with parameter names to avoid further cluttering an already busy display.

Since a single, scaled latent variable is used to induce dependency, this approach is less flexible than one in which in which response-specific latent variables are used. In particular, this parameterisation implies a linear dependency between response variables.

The estimated factor loadings offer information about the direction of any dependencies (after controlling for covariates), but do not directly offer information about the magnitude. They can be used to calculate so-called tetrachoric or polychoric correlation coefficients however (Skrondal and Rabe-Hesketh, 2004). In practice, we find that switching to the shared parameter model offers no advantage in terms of improving convergence. Autocorrelation remains high and effective sample sizes after running the chain for several thousand iterations is low.

Removing the latent dependency between the number of oocytes submodel and downstream responses

An alternative approach we might consider is to remove the dependency between the latent variable in the 'number of oocytes' submodel and the latent variables from the downstream submodels. So that the underlying variance-covariance matrix for the latent variables in the model becomes:

$$\Theta = \begin{bmatrix} \theta_O^2 & 0 & 0 & 0 & 0 \\ 0 & \theta_N^2 & \eta_5\theta_N\theta_E & \eta_6\theta_N\theta_F & \eta_7\theta_N \\ 0 & \eta_5\theta_E\theta_N & \theta_E^2 & \eta_8\theta_E\theta_F & \eta_9\theta_E \\ 0 & \eta_6\theta_F\theta_N & \eta_8\theta_F\theta_E & \theta_F^2 & \eta_{10}\theta_F \\ 0 & \eta_7\theta_N & \eta_9\theta_E & \eta_{10}\theta_F & 1 \end{bmatrix}$$

The improvement in mixing compared to previous versions of the model is considerable. We run three chains for 2000 iterations, discarding the first half as burn in. We observe excellent convergence on the basis of Gelman-Rubin statistics and traceplots, as well as much larger effective sample sizes compared to previous fits based on longer chains.

An autocorrelation plot of the elements of the correlation matrix (Figure 33) shows the drastic reduction compared to the initial fit (as shown in Figure 31). This improvement comes at a cost however, since severing the correlation between the number of oocytes and the downstream responses corresponds to treating number of oocytes as an exogenous variable. Our estimates of the effects of number of oocytes on downstream responses from this model therefore rest on an assumption that there is no unmeasured confounding.

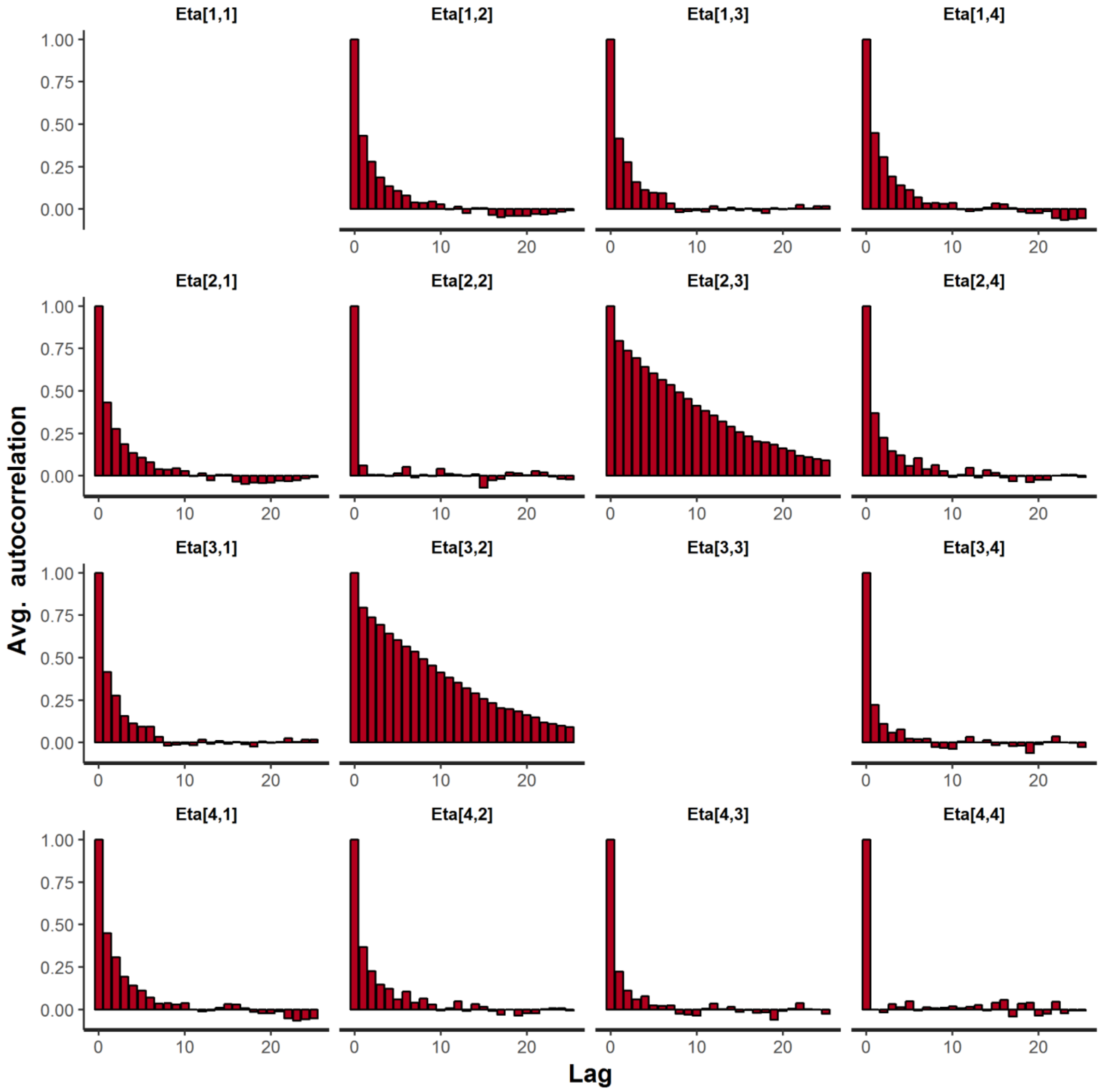


Figure 33: Autocorrelation plot showing the correlation between parameter draws at lag x, for each element of the coefficient matrix in the shared latent variable reparametrisation of a joint model of ovarian stimulation, embryo culture and embryo transfer which includes the variable 'number of oocytes', representing response to ovarian stimulation, as a covariate in the submodels for the downstream stages of treatment.

Using instrumental variables to identify the model

One reason for poor mixing of the chains might relate to identifiability. Although the parameters of the model discussed here are theoretically identified, they may only be weakly identified by the data. Identification may be particularly difficult due to the fact that the same covariates are included in the various submodels (with the exception of the variable number of oocytes, which of course does not appear as a covariate in the model of itself). It may be possible to improve the identification of the model parameters by including covariables which appear in no more than one of the submodels. Such covariables can be considered instrumental variables; we suppose that they have an effect upon the response 'number of oocytes', but do not have any effect on downstream responses by any other route. Instrumental variables have been used to improve identification in joint models including responses as covariables (for example, the endogenous switching models presented by (Kenkel and Terza, 2001, Terza, 1998, 2000, Xie, 2000).

This prompts the question of which variables would be appropriate in this role in the present example. One possibility would be to restrict the terms relating to attempt number to the stimulation stage submodel. The rationale here is that the dose of drug (gonadotropin) used for ovarian stimulation is likely to be modified for subsequent treatment attempts according to the stimulation response (number of oocytes) in the previous cycle. It could be hypothesised that changes to the dose of the drug would impact the quantity of oocytes obtained from stimulation, but would not influence the quality of embryos or the likelihood of transfer success, other than by virtue of the fact that more or less eggs were available for fertilisation, and that this would increase or decrease the number of embryos from which to select the best for transfer. In fact, this may not be so plausible because it has been suggested (if not demonstrated) that increased doses of drug may in fact make the uterine environment less hospitable to transferred embryos (eg: Maheshwari and Bhattacharya, 2013); this is the question we attempt to answer in Journal Article 6 (Chapter 11).

In the embryo submodels, an obvious candidate to play the role of instrumental variable is the method of fertilization (mixing or injecting with sperm). This is because it is difficult to imagine how the fertilization could influence downstream outcomes such as the result of embryo transfer, other than by way of the embryos produced. It is possible that that method of fertilization and transfer outcomes could share unmeasured predictive variables, such as the cause of infertility (which might cause the clinicians to prefer one fertilization protocol over the other, and might also impact on transfer success). In that case, the so-called 'back door criterion' would be violated, and method of fertilization would not be a valid instrument (Emsley, et al., 2010). This would not be problematic however, since we are not relying on our 'instruments' for causal identification; the correlated latent variables in the model serve this function by representing the dependence between response variables induced by unmeasured confounding. Instead, we use these (possibly invalid) instrumental variables as a way to fit the model.

We investigate the instrumental variable principle here by setting the parameters relating to attempt number to zero in the downstream response models. We also add the binary 'method of fertilization' indicator variable to all three embryo quality submodels. Figure 34 shows autocorrelation plots for the latent correlation coefficients from this model. Although the autocorrelation is still substantial, it decays more rapidly than in the endogenous response model without instrumental variables, and reaches zero at approximately 60 lags.

As a strategy to improve convergence, the introduction of instrumental variables is less effective, but nonetheless preferable, compared to fixing the latent correlation between oocytes and the downstream responses to be zero; if the latter strategy is adopted, the benefits of fitting an endogenous response model are sacrificed.

7.5 A note on our use of causal language in relation to endogenous response models

In the preceding sections, we have referred to 'effects' and 'causal effects' of response variables on downstream responses. In the ubiquitous counterfactual formulation of causal inference, it is incoherent to speak of effects of variables that are not directly manipulable (Greenland, 2017). As such, in that framework, it would not be possible to speak of effects of age or of the number of oocytes on other variables. Some authors consider this to a *reductio ad absurdum* of the counterfactual framework; since it is clearly

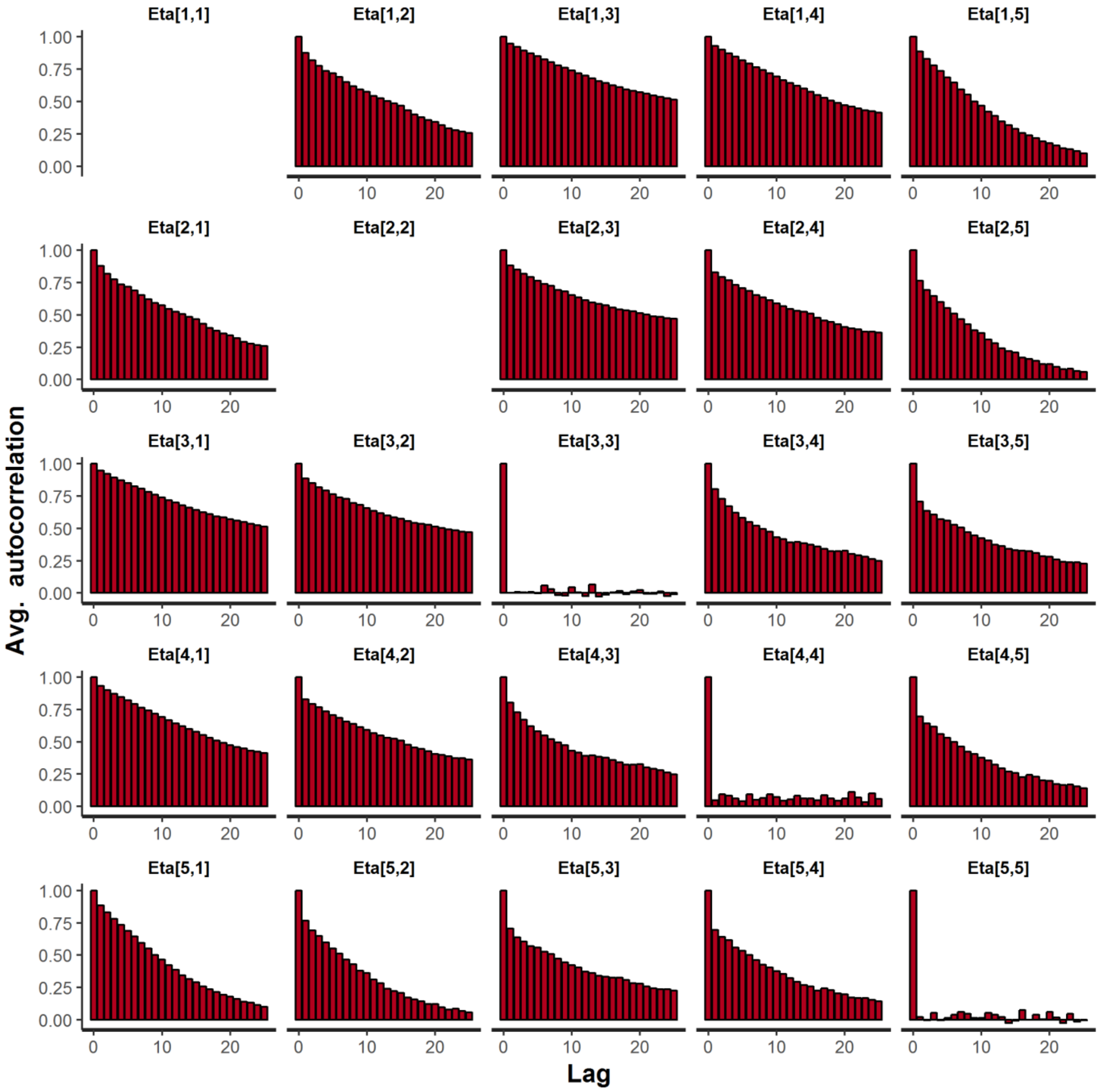


Figure 34: Autocorrelation plot showing the average correlation between parameter draws at lag x, for each element of the coefficient matrix in the joint model of ovarian stimulation, embryo culture and embryo transfer which includes the variable 'number of oocytes', representing response to ovarian stimulation, as a covariate in the submodels for the downstream stages of treatment. The instrumental variables 'attempt number' and 'method of fertilization' have been included in the ovarian stimulation and three embryo culture submodels.

reasonable (they argue) to speak of an effect of age on patient outcomes, the counterfactual framework should be rejected (Krieger and Davey Smith, 2016). The objection to our usage might have some merit however. For example, we might speak of the 'direct effect' of embryo quality on the success of the embryo transfer procedure, for fixed values of upstream variables. However, since embryo quality is causally dependent upon upstream variables, it doesn't obviously make sense to speak of increasing it while holding those variables fixed. Indeed, there is no actual (or possibly even hypothetical) referent of the phrase ‘increasing embryo quality’, since this is not something which we are able to manipulate directly. We might, however, be able to manipulate interventions earlier in the IVF cycle which will improve embryo quality, and thereby increase the success probability of the embryo transfer. In this case, we should speak of the effects of the intervention on embryo transfer *being mediated through* embryo quality (Emsley, et al., 2010).

Rather than weigh in on this debate, we continue to use the term ‘effect’ for both manipulable and non-manipulable variables in our exposition, noting that various proposals have been made to legitimise this usage (Pearl, 2011, VanderWeele and Hernán, 2012). The models could be redescribed using the language of counterfactuals and mediation for nonlinear models (Pearl, 2011), although this would not obviously offer any practical advantage when using these models to answer clinical questions.

7.6 **Moving from toy examples to real applications: adding submodels**

We have described the models presented in this chapter as simplistic, since they exclude important components of the IVF cycle and are unlikely to be useful for real applications. This prompts the question of which additional response variables should be introduced in order to make the model useful for the purpose of tackling real problems. A useful thought experiment in this regard is to imagine a patient going through IVF, and to consider to what extent our model could predict the patient’s responses at each stage of the process. While the exact submodels and covariates to include will depend on the particular research question under consideration (and Journal Article 6, Chapter 11, provides an example), we can use this exercise to identify important milestones in the IVF cycle which we anticipate should feature in most applications. Note that there is no contradiction in using a thought experiment based on prediction to inform an explanatory model; the responses we would want to predict in an IVF cycle are the very same as those we would like to subject to mechanistic inquiry.

A patient undergoing IVF usually begins with ovarian stimulation. We could predict the number of eggs that will be obtained on the basis of baseline characteristics and treatment variables, using the Poisson formulation featured in the toy examples presented so far. The patient's eggs will then be then fertilized with sperm, to produce embryos. Our examples to date omit this stage entirely, jumping straight from egg collection to the quality of any embryos produced. As a result, we would be unable to predict how many embryos the patient would be produced from the patient's eggs. Moreover, since our embryo quality submodels are specified at the level of individual embryos, we would be unable to predict the quality of the patient's embryos without first making an arbitrary decision as to how many embryos we should predict for. These considerations suggest that a submodel relating to the fertilization of eggs, with the response variable 'number of embryos obtained', should be included in the model. The model could then predict how many of the patient's eggs will be successfully fertilized, and the quality of each of the resulting embryos. To this end, we introduce an overdispersed Poisson submodel for number of embryos in Journal Article 5 (Chapter 10), including the number of eggs obtained as an offset term.

Following the fertilization and culture of embryos, some of these will be transferred to the patient's uterus. The number of embryos transferred is thought to be an important predictor of transfer outcome, and in most applications we would probably include this as a predictor in our live birth submodel. However, the number of embryos transferred depends in part on upstream responses, including the number and quality of embryos available following stimulation, fertilization and embryo culture. Patient characteristics such as age also determine how many embryos should be transferred, and are often explicitly included in standard operating procedures. Accordingly, there is a strong case for including the number of embryos transferred as an endogenous response in the model. We introduce a binary response variable denoting whether one or two embryos are transferred in Journal Article 5 (Chapter 10), and model this using a latent variable probit regression submodel, as we have for live birth event in the examples discussed so far.

Our live birth event submodel could be used to predict whether or not the procedure will be successful, once the embryos have been transferred. This might be too coarse for some applications however, because the procedure could fail for different reasons. For example, the transferred embryos could fail to implant in the uterine wall. Alternatively, they could implant, but the patient could suffer a subsequent miscarriage. In Journal

Article 6 (Chapter 11), we distinguish between embryo implantation, and live birth event conditional on embryo implantation, both modelled as binary variables using latent probit submodels. This allows us to consider whether increasing gonadotrophin dose has deleterious consequences on embryo implantation and on foetal development in the uterus separately.

7.7 Adjusting for unmeasured confounding: proof of principle using simulated data

While they have featured in the econometrics literature for over forty years (Heckman, 1976), methods to adjust for unmeasured confounding variables are not particularly common in the medical literature (although they are not completely unknown, eg: Streeter, et al., 2017). Understandably, the idea that adjusting for unmeasured variables is even possible has proven troublesome for some discussants of the present work. Accordingly, we present here a small simulation study intended to illustrate how the endogenous response method can potentially mitigate unmeasured confounding.

Given the time they take to fit, it isn't feasible to carry out simulations using models of the full IVF cycle. Here, we emulate a scenario arising in Journal Article 6 (Chapter 11). In that article, we attempt to assess the effects of log(dose) of ovarian stimulation drugs on outcomes occurring at different stages of the cycle. This includes the effect on the number of oocytes obtained following the stimulation period. There is known to be counfounding due to the fact that the dose administered depends on the anticipated response. The characteristics used for dose selection are not recorded in the database used for analysis, however, so we have unmeasured confounding between dose and number of oocytes.

7.7.1. Data generating model

We simulate datasets corresponding to this scenario. Each contained 2000 IVF stimulation cycles, which resembles the size of the datasets we consider in the thesis. In a single iteration, for each cycle, we simulate an instrumental variable, a covariate, and a confounder, all from Normal(0,1) distributions. We simulate log(dose) from a Normal distribution with standard deviation 1.5 and mean equal to 7 + 2*INSTRUMENT + 0.7*COVARIATE + 0.7*CONFOUNDER. We then simulate number of eggs from a Poisson distribution, with log(mean) equal to log(5) + 0.3*LOG(DOSE) + 0.05*COVARIATE + 0.3*CONFOUNDER + 0.3*CONFOUNDER^2. The values of the coefficients were chosen so that the confounding would be of similar magnitude to the treatment effect. We included the quadratic term for the confounder to represent misspecification of the latent variable distribution in our analysis model.

7.7.2. Analysis models

Our target is the effect of log(dose) on number of oocytes (equal to 0.3 in our data generating model). We analyse the simulated datasets using a joint model of log(dose) and number of oocytes, which includes a linear regression and a Poisson submodel. As in this chapter and elsewhere, we join the submodels by specifying a bivariate Normal distribution for the latent variables appearing in each (the residual in the log(dose) submodel, and an overdispersion term in the oocytes submodel: see Journal Articles 5 and 6, Chapters 10 and 11). In the linear predictor for the log(dose) submodel, we include an intercept, the instrumental variable, and the covariate. In the oocytes submodel, we include an intercept, the covariate, and log(dose), making this an endogenous response model. For comparison, we also estimate the effect of dose on oocytes by fitting a univariate overdispersed Poisson regression to the number of oocytes, again with an intercept, the covariate, and log(dose).

7.7.3. Results of the simulation

We ran the simulation for 40 iterations. We stress that this is not intended as an evaluation of the properties of the endogenous response method, which would require data generated under different scenarios and many more iterations for each. We simply hope to provide some reassurance to any reader who is wary about the prospect of adjusting for unmeasured confounding. Results are presented in Table 12.

In the scenario considered here, the endogenous response model outperformed simple outcome regression with respect to bias, efficiency, and coverage. Bias was 45 times greater using outcome regression compared to using the joint model. The maximum error in estimation using the joint model was 0.02, compared to 0.24 using outcome regression. There was also a substantial efficiency gain from jointly modelling the responses. Finally, coverage was hopeless using outcome regression (only 5% of 95% CIs contained the true value). By contrast 95% of the 95% intervals obtained from the joint model contained the true value.

	Endogenous Response Model	Poisson, with dose as covariate (outcome regression)
Bias	0.002	0.09
Mean SE	0.006	0.02
Coverage of 95% CI	95%	5%

Table 12: Results of small simulation study, including a Normal response and a Poisson response, with unmeasured confounding.

7.8 Summary of Chapter 7.

In this chapter we have introduced two multistage modelling approaches for mechanistic analysis of IVF data. Both approaches involve the specification of submodels for the responses included in the model. Each of these are fitted conditional on the patient not having outright failure upto that stage. Dependency can be accommodated and quantified by introducing correlated latent variables in the response-specific submodels. However, the latent correlation coefficients obtained from this approach are not obviously interpretable, or of practical use. An extension of this approach, where we include response variables as covariates in the submodels for downstream responses, offers an advantage in this regard, by providing adjusted estimates of effect on an interpretable scale. It is essential that we retain the correlated latent variable structure when doing so, since this allows for unmeasured confounding between response variables. We call models of the second type endogenous response models. We found that estimating the parameters of endogenous response models is difficult in practice, requiring long runs of the sampling algorithm in order to obtain a sample from the posterior distribution of satisfactory size. The introduction of instrumental variables in some of the submodels alleviated, but did not completely eliminate, the convergence issues. In a simple simulation exercise, we illustrated how endogenous response models can adjust for unmeasured confounding.

The approaches presented in this chapter all assume that missing data due to drop out can be explained as MAR, given the covariates and response variables included in the model. In the next section, we discuss the matter of missing data due to drop out in more detail, and outline strategies for relaxing the MAR assumption.

7.9 References for Chapter 7.

Diggle P, Farewell D, Henderson R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *J R Stat Soc Ser C Appl Stat* 2007;56: 499-529.

Dunson DB. Bayesian latent variable models for clustered mixed outcomes. *J R Stat Soc Series B Stat Methodol* 2000;62: 355-366.

Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* 2003;59: 521-530.

Dunson DB, Herring AH. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* 2005;6: 11-25.

Emsley R, Dunn G, White IR. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Stat Methods Med Res* 2010;19: 237-270.

Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *J Res Educ Eff* 2012;5: 189-211.

Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci* 1992: 457-472.

Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. *Stat Model* 2009;9: 173-197.

Greenland S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *Eur J Epidemiol* 2017.

Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized Regression Coefficients - a Further Critique and Review of Some Alternatives. *Epidemiology* 1991;2: 387-392.

Gueorguieva R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat Model* 2001;1: 177-193.

Heckman JJ. Common Structure of Statistical-Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Ann Econ Soc Meas* 1976;5: 475-492.

Kenkel DS, Terza JV. The effect of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. *J Appl Econom* 2001;16: 165-184.

Krieger N, Davey Smith G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemio* 2016.

Maheshwari A, Bhattacharya S. Elective frozen replacement cycles for all: ready for prime time? *Hum Reprod* 2013;28: 6-9.

McConnell S, Stuart EA, Devaney B. The truncation-by-death problem - What to do in an experimental evaluation when the outcome is not always defined. *Evaluation Rev* 2008;32: 157-186.

McCulloch C. Joint modelling of mixed outcome types using latent variables. *Stat Methods Med Res* 2008;17: 53-73.

Pearl J. Direct and indirect effects. *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* 2001: 411-420.

Pearl J. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. 2011. CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.

Rubin DB. Inference and Missing Data. *Biometrika* 1976;63: 581-590.

Skrondal A, Rabe-Hesketh S. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, 2004. Crc Press.

Stan Development Team. Stan Modeling Language: User's Guide and Reference Manual. 2017.

Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer Science & Business Media; 2008.

Steele F, Washbrook E. Discrete-time Event History Analysis. In Centre for Multilevel Modelling UoB (ed). 2013.

Streeter AJ, Lin NX, Crathorne L, Haasova M, Hyde C, Melzer D, Henley WE. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. J Clin Epidemiol 2017 Apr 28.

Terza JV. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *J Econometrics* 1998;84: 129-154.

VanderWeele TJ, Hernán MA. Causal effects and natural laws: towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex, 2012. Wiley Online Library.

Xie Y. Endogenous Switching Regression Models in *Into Adulthood: A Study of the Effects of Head Start*. 2000, pp.169-178.

Chapter 8. Accounting for drop out in the IVF model

As patients progress through the stages of the fresh IVF cycle, a nontrivial number drop out prior to egg collection and embryo transfer. In Journal Article 3 (Chapter 5), we referred to the 2014 National Summary Report of the Society for Assisted Reproductive Technology (SART) and noted that 9247 of 102,982 (9%) cycles were terminated prior to the egg collection stage. Of the 93,730 egg collections that did occur, 7188 (8%) did not result in a transfer procedure due to a lack of embryos. This complicates analysis of IVF data. Although it should be straightforward to obtain an intention-to-treat estimate of the effect of an intervention on the probability of a live birth event (since we can define this outcome to be a failure in patients who drop out or fail earlier stages of treatment), in practice researchers struggle with the implications of attrition. For example, in a review of outcome measures in IVF RCTs between 2013 and 2014, 87 distinct denominators were used, with many of these representing subgroups of patients who achieved a certain stage in the treatment (such as egg collection or embryo transfer) (Journal Article 2, Chapter 4). Fifty-eight per cent of reported live birth rates were calculated in a post-randomisation cohort of patients. In relation to our goal of developing multistage mechanistic models, there is also the matter of how to treat procedural responses such as the number of oocytes obtained or embryo quality given that they are unobserved or undefined for patients who exit treatment in the preliminary stages.

This presentation of the IVF cycle as a sequence of observations on a patient, which may not be fully realised as a result of attrition, clearly evokes a longitudinal data analysis framework (Diggle, 2002). Consequently, we turn to the literature relating to missing data in repeated measures contexts in order to investigate how drop out could be incorporated into our modelling framework. In the following, we briefly review the main approaches to longitudinal analysis with missing data and discuss the implications for analysis of IVF data with drop out, in light of the fact that the repeated measurements in the present context comprise a sequence of mixed outcome types occurring at different levels of a multilevel hierarchy.

8.1 What is the objective of analysis when missing data due to drop out or censoring are present?

Diggle and colleagues (2007) have noted that the objective of missing data analysis is often left obscure in practice. When an observation is missing due to drop out, they distinguish between two possible targets. The first is the extant (but unobserved) value of the response. In principle, we could obtain this given sufficient access to the patient. The second is the counterfactual value of the response that would have been obtained had the patient not dropped out of the study. In the following we refer to the former as 'extant-missing' and the latter as 'counterfactual-missing'. These two are conceptually distinct and their values may or may not coincide. The authors note that the appropriate approach to handling missing data arising from drop-out partially depends on which of these quantities is of interest. If the appropriate method of analysis should be governed by the scientific objective of a particular study, it follows that no single approach will be generally applicable to IVF data. Instead, we must remain open to the prospect that different approaches may be appropriate in different settings. An example can be found in our investigation of personalised ovarian stimulation (Journal Article 4, Chapter 9). There, we were interested in the effect of increasing the dose of ovarian stimulation drug on the yield of oocytes collected. Oocyte yields are not observed in patients who have their treatments 'cancelled' partway through on the basis of ultrasound monitoring. This may be due to either an anticipated poor or excessive response. In our analysis, we imputed these values using an intermediate response variable (counts of egg-releasing follicles) and additional patient and treatment covariables. We decided that this was appropriate based on our interest in the dose-response relationship between drug and outcome; we were interested in the counterfactual-missing values that would have been observed had stimulation not been cancelled in these cycles. However, had our objective been to make a more pragmatic evaluation of the impact of different dosing regimens on patient outcomes, it may have been more appropriate to define all egg counts in cancelled cycles to be zero, regardless of whether the cancellation was due to anticipated poor or to anticipated hyperresponse. It is debatable whether this would meet the description of 'extant-missing' data above; although a value of zero does correspond to the actual number of eggs collected for these cycles, there is no additional 'true' value of the response hidden away. Instead, it may turn out to be conceptually appropriate (as well as intuitively appealing) to declare these egg yields to be strictly undefined, given that no egg collections have taken place. We return to this point later. For now, we note that the multistage nature of IVF means that decisions relating to the target of inference may differ not only from study to study, but also for the different response variables being modelled within a study.

8.2 Mechanisms of missingness

So the objective of the data analysis is one consideration when determining an appropriate statistical strategy. A second is the process giving rise to the missing data. Following the classifications introduced by Rubin (1976), we consider drop-out to be missing completely at random (MCAR) if the probability that an observation is missing is unrelated to both the observed and unobserved data (including both covariates and response variables), to be missing at random (MAR) if the probability of missingness is related to the observed data but, once this has been conditioned on, unrelated to the unobserved data[4] and missing not at random (MNAR) in the event that the probability of missingness is related to the unobserved data and remains so after conditioning on observed data. In practice, the missing data mechanism is strictly unknowable and must be the subject of assumptions and sensitivity analyses. Immediately then, we can see that whichever mechanism we deem to be the most plausible may depend on several factors, including which particular variables are incomplete and which are available to the analyst as conditioning variables.

A further consideration is that, since IVF comprises a sequence of distinct stages, the drop out mechanism may realistically vary throughout treatment. Consequently, in addition to the possible targets of inference at each stage of treatment, it is worth reflecting on the plausible mechanisms underlying the main causes of drop out from the IVF cycle.

[4] A distinction is sometimes made between the case where data are MAR given covariate data and that where they are MAR given the observed longitudinal response data. We do not require this distinction in the present discussion, and refer to the case where missing data are ignorable given whatever data are included in the model, as MAR.

8.2.1. What is (are) the mechanism(s) of missingness in IVF?

Our methodological focus in this thesis lies in the development of models for the multistage IVF cycle. As a result of this and due to a lack of multiple cycle datasets, we focus on drop out during the sequence of interventions delivered within a given cycle in the following discussion. The question of why patients do not continue courses of treatment involving repeated treatment cycles has been discussed elsewhere (Verberg, et al., 2008) and several methodological approaches for handling this have been proposed (Hogan and Scharfstein, 2006, Soullier, et al., 2008).

Cancellation of ovarian stimulation

We briefly discussed above (1.1.) the case where the IVF cycle may be cancelled due to anticipation of an undesirable outcome to ovarian stimulation, and gave an example in which the outcome number of oocytes was imputed in the analysis. As such, we set the counterfactual-missing values as the target of inference and considered these to be MAR given the covariables in the imputation model. This was probably plausible since information on a reasonable surrogate outcome, the number of egg-releasing follicles observed on ultrasound, was available for this purpose. We also noted above that a pragmatic assessment of predictors of number of oocytes might set the responses of cancelled cycles to be zero, since these cycles yield no eggs for the fresh transfer (in the case of hyperresponse, the eggs may or may not all be frozen for transfer in a subsequent cycle). A third possibility would be to say that the outcome is strictly undefined for these cycles, given that no egg collection has taken place. A proponent of this view would declare this to be an instance of 'truncation-by-death' (Rubin, 2006, Zhang and Rubin, 2003), with another example being the impossibility of defining marriage quality in individuals who have divorced (McConnell, et al., 2008). If so, it doesn't obviously make sense to speak of the missingness mechanism, since there are no underlying missing values (either of the extant or counterfactual variety) to be explained.

However, some consideration of the process giving rise to the outcome allows us to challenge this position. The administration of gonadotropins to the patient causes follicles to be recruited, each of which contains an immature oocyte. Recruited follicles will grow and produce mature eggs. This process is monitored by ultrasound, and when the follicles are ready, a trigger hormone is administered permitting the harvesting of these eggs. If too few or too many follicles are present, the cycle may be cancelled on the grounds of futility in the former case and safety in the latter. The implication is that, had the stimulation not been cancelled, the eggs could have been collected from the follicles. Accordingly, it may be reasonable to consider the value of number of oocytes to be censored (in the sense that a value exists but is unknown) rather than truncated (and hence, undefined). If so, counts of the numbers of large follicles could be used to impute or predict the number of oocytes under the MAR assumption. The prediction will not be perfect however, since an egg may not be retrieved from some follicles and there appears to be variation in the number collected corresponding to who carries out the procedure (Journal Article 4, Chapter 9). Of course, the surgeon conducting the procedure cannot be included in the imputation model since for cancelled cycles no procedure has been performed. Even if a count of follicles is available therefore, residual confounding might guarantee that the missing values of number of oocytes, if we can consider them to be censored rather than undefined, are really MNAR.

No embryos for transfer

A cycle may have to be abandoned because there are no embryos available for transfer to the patient. This could be because there were no eggs following ovarian stimulation, or because there were no eggs which were fertilised and subsequently developed into usable embryos. In the event that no eggs were collected, then neither the number of embryos obtained nor the quality of the patient’s embryos are observed or defined. This would appear to be a more cut and dry example of truncation-by-death than number of oocytes in cancelled cycles. We could try to conceive of counterfactual-missing values for these response variables (eg: the number of embryos that would have been obtained, and their quality, had eggs been collected) but it is impossible to predict or impute these quantities under an MAR assumption without making the strong assumption that the effects of covariates on embryo responses in patients who have eggs would be the same in the patients who don’t have eggs were they to (counterfactually) have some. One proposal would be to investigate (and perhaps relax) this assumption using data relating to repeated cycles on the same participants, where some (but not all) of these cycles have been cancelled prior to egg collection. In reality, this would not overcome the issue so much as relocate it to the level of the cycle; we would still have to assume that the relationships between variables and embryo responses were the same for any given patient in cycles where they did and did not have cancellation prior to egg collection. Even if we allowed the missing values to be MNAR, we would still have to make assumptions about the latent variable distribution in the drop-outs (see section 1.3).

If eggs are fertilised, but do not develop into usable embryos (for example, they do not survive to the transfer day, or are of insufficient quality for transfer), then both number of embryos and measures of quality may be available, provided that these are measured sufficiently early following fertilisation (on day 2 for example).

Implantation failure and pregnancy loss

Embryos transferred to the patient may fail to implant in the uterine wall, precluding further development. As such, the patient does not become pregnant in this scenario. It would usually be inappropriate to define the clinical outcome of the cycle as missing, rather than as a failure. For some research questions however, we may wish to distinguish between an embryo failing to implant (resulting in no pregnancy) and the implantation not being sustained (resulting in pregnancy loss). We make this distinction in Journal Article 6 (Chapter 11), where our interest is to establish the effect of ovarian stimulation on embryo implantation and birth. One can also imagine specific situations where it would be reasonable to think about the likelihood that a pregnancy would have progressed to a live birth had an embryo implanted (for example, in a trial of an intervention designed to reduce miscarriage). The prevailing approaches to this in IVF RCTs are to either calculate miscarriage rates per woman randomised (such that an intervention that results in no pregnancies would result in no miscarriages, (eg: Ferraretti, et al., 2014, Gao, et al., 2013, Revelli, et al., 2014) or per pregnancy (discarding the benefits of randomisation, eg: Alviggi, et al. (2013) Check, et al. (2013, Friedler, et al. (2013)). Similar considerations relate to participants who did not make it as far as the transfer stage. Is it reasonable to ask what would have happened in a transfer had the patient made it that far? These appear to be further cases of truncation-by-death. We note that any approach that attempts to answer these hypothetical questions will be forced to rest on strong assumptions concerning similarities between those who did and did not have an embryo implantation or a transfer procedure.

8.3 Statistical methods for incorporating drop out in longitudinal data

We briefly review several approaches for modelling drop out in longitudinal data settings. In each case, we discuss the features of the approach that make it more or less suitable for modelling dropout in the fresh IVF cycle.

8.3.1. Diggle-Kenward model

Diggle and Kenward (1994) presented a model for drop out in longitudinal settings. We follow the exposition of Skrondal and Rabe-Hesketh (2004) here. To illustrate the model, consider the archetypal longitudinal data analysis setting with repeated measurements of a single continuous outcome *y* at time *i* for unit *j* :

$$y_{ij} = \mathbf{X}_{\mathrm{ij}}\boldsymbol{\beta} + u_j + \epsilon_{ij}$$

where u_j and ϵ_{ij} are Gaussian residuals at the level of the participant and of the observation respectively, with zero means, and variances ($\sigma_u{}^2$ and $\sigma_e{}^2$, say), and independence from commensurate residual terms unless their subscripts coincide. $\boldsymbol{\beta}$ is a vector of regression coefficients and $\mathbf{X}_{\mathrm{ij}}$ is a row-vector of covariates. We define a dropout model;

$$logit\{P(d_{ij} = 1|\boldsymbol{Y_j}, \alpha_0, \alpha_1, \alpha_2)\} = \alpha_0 + \alpha_1 y_{ij} + \alpha_2 y_{i-1,j}$$

Where $\boldsymbol{Y_j}$ is the participant's response-vector, and $d_{ij} = 1$ if y_{ij} is unobserved due to dropout at time *i*. If d_{ij} does equal 1, then we replace y_{ij} in the linear predictor with a latent variable y_{ij}^*. Once drop-out has occurred, we do not model d_{ij} for subsequent timepoints. This approach allows the probability of dropout to depend on the current (possibly unobserved) response, and the previous response (or responses, since we could include earlier measurements in the linear predictor if desired). If α_1 is equal to zero, drop-out is MAR. If α_2 is also equal to zero, then drop-out is MCAR (Diggle, et al., 2007). Otherwise, this can be viewed as a MNAR analysis, where the likelihood that a response is observed depends on both its value and on the values of upstream responses. Diggle and colleagues (2007) note that a tacit assumption of this approach is that the counterfactual-missing and extant-missing values are identical. In addition, valid estimation in this framework depends on the correct model specification for both the response vector $\boldsymbol{Y_j}$

and the drop-out vector $\boldsymbol{D_j}$. Consequently, this approach may be best employed as a sensitivity analysis of the MAR assumption (Skrondal and Rabe-Hesketh, 2004).

Adaptability to the present setting

A modification of the Diggle-Kenward model could be incorporated into a multistage model of the IVF cycle. Instead of the Normally distributed continuous outcome variable posited in the example above, we would include the sequence of mixed responses arising from each stage of treatment. The probability of dropout at each stage could be modelled using a logistic or probit regression. For a MNAR analysis, we could include the response at the present stage as a covariate in the logistic regression, together with the upstream responses. For a MAR analysis, only the upstream responses would be included. This approach would allow the dropout model to vary across stages, by including different covariates at each stage. Accordingly, we could model the dropout process at one stage as MNAR (for example), while modelling the remainder as MAR. Given our discussion of the multiple mechanisms of missingness in the IVF cycle above, this aspect of the (modified) Kenward-Diggle approach is attractive. In practice however, this modification could not be implemented in Stan. The algorithm does not support discrete parameters (not to be confused with regression parameters corresponding to discrete variables). Consequently, we would not be able to model y_{ij}^{*} for our discrete outcomes (number of eggs, embryo evenness etc).

8.3.2. Discrete time-to-event logistic submodel with correlated random effects

Let's return to our conventional mixed model for a longitudinal continuous response variable, with model at time *i* for unit *j* :

$$y_{ij} = \mathbf{X}_{\mathrm{ij}}\boldsymbol{\beta} + u_j + \ \epsilon_{ij}$$

With Gaussian random effects and level 1 residuals, covariates and regression coefficients as described above. We could specify a discrete time-to-event model for the dropout process (Steele, et al., 2009):

$$logit\{P\left(d_{ij} = 1 \middle| \alpha_i, c_{ij}, \boldsymbol{\gamma}_i, v_{\boldsymbol{j}}\right)\} = \alpha_i + \mathrm{c}_{\mathrm{ij}}\, \boldsymbol{\gamma}_i + v_j$$

which now includes a row-vector of covariates c_{ij} , a latent variable v_j that is unique to the unit and a vector of regression coefficients $\boldsymbol{\gamma}_i$. This is a model for the hazard of dropping out. The intercept term α_i now bears an *i* subscript. This can be interpreted as a step function denoting that the baseline hazard is time-varying.

We can specify a bivariate Normal distribution for the latent variables (*u*,*v*) ~ N(0, ***V***) so that dependency between the drop-out process and the longitudinal responses are represented through the latent covariance matrix:

$$\boldsymbol{V} = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}$$

We make the assumption that the responses y_{ij} and drop-out indicators d_{ij} are independent given the latent variables. The correlated latent variable approach is discussed by Diggle, et al. (2007), who note that it is a MNAR approach, due to the fact that the conditional distribution of the latent variables depends on the values of the responses at all timepoints, including counterfactual-missing values.

Adaptability to the present setting

This approach can be viewed as an extension to the correlated latent variables methods we have used to jointly model the stages of the fresh IVF cycle in previous sections, with a submodel for the dropout process linked to the response portion of the model through random effects. As such, it could be incorporated into the current modelling framework. Different covariates can be included at each stage. A major limitation of the approach however is the need to include a single latent variable v_j common to all stages of the cycle. Given our discussion of potentially varying drop out mechanisms in the different stages of the cycle above (8.2.1), this is likely to make this approach too restrictive for our purposes. The assumption of a common random effect across stages could be relaxed were we to use multicycle data (that is, where each woman contributes multiple fresh IVF cycles). In that scenario, our dropout submodel would be similar to the model described by Maity, et al. (2014). Without repeated cycles however, stage-specific random effects are not identified in a logistic dropout submodel.

8.3.3. Sequential probit submodel with correlated random effects

An alternative to the discrete time-to-event model described above is the sequential probit model (Steele and Washbrook, 2013, Waelbroeck, 2005). This can be used when a subject must successfully pass through a series of stages in order to enter the next in the sequence. We define a binary variable b_{ij} which equals 1 if participant *j* achieves timepoint *i* + 1 and 0 if they do not progress beyond stage *i*. We define a latent variable b_{ij}^{*}, such that b_{ij}^{*} is negative if b_{ij} = 0 and is greater than or equal to zero otherwise. We then specify a linear regression submodel for the latent variable:

$$b_{ij}^{*} = \mathbf{C}_{\mathrm{ij}}\, \boldsymbol{\gamma}_{i} + \mathrm{w}_{\mathrm{ij}}$$

which includes a row-vector of covariates $\boldsymbol{C}_{\mathrm{ij}}$, a vector of timepoint-specific regression coefficients $\boldsymbol{\gamma}_{i}$ and a Gaussian residual term w_{ij} with mean zero and variance of 1. We allow the residual term to be correlated with the latent variables from the longitudinal response model, and estimate the covariance parameters from the data.

Adaptability to the present setting

The sequential probit model for the dropout process can be incorporated into our existing joint modelling framework. It is flexible compared to the discrete time-to-event logistic approach detailed in 8.3.2 as it permits the identification of stage-specific residual terms. When embedded in a larger joint model, this then allows the correlation between each stage-specific response and stage-specific dropout probabilities to be estimated. An example is a multiprocess model including a sequential probit submodel was developed by Steele, et al. (2009), who jointly modelled the stage of education reached by a child with the dissolution of the mother’s marriage. They linked the sequential probit model and the time to marriage dissolution submodel by simultaneously including correlated mother-level random effects in the two submodels and including prior marriage dissolution outcomes as endogenous covariates in the education submodel. Returning to the case of IVF, dependency between drop out and treatment responses could possibly be incorporated both through correlated latent variables and by including upstream responses as covariates. As with the discrete time logistic model, this is a MNAR approach, as the counterfactual-missing outcomes contribute to the conditional random effects distribution. This may be problematic if we do not wish to represent dropout at certain stages by a MNAR mechanism. In particular, we return to the concerns, raised in section (8.2.1), that the values should be considered undefined. If so, implicit assumptions about the latent variables (specifically, that the latent variable distribution will not vary depending on whether or not a patient's multivariate response vector is fully or only partially observed), are not obviously coherent (McConnell, et al., 2008).

8.4 Application of the sequential probit dropout model

Despite our unresolved reservations about the necessity, or even suitability, of incorporating an informative dropout model, we fitted an endogenous response model extended to include a sequential probit submodel for the dropout process. The response model is a variation of that described in detail in Journal Article 5 (Chapter 10). It includes the following outcome variables (regression submodels): number of oocytes (Poisson), fertilisation rate (Poisson with number of oocytes as an offset); embryo evenness and fragmentation (two-level cumulative logit); double embryo transfer (probit) and live birth event (probit). In this simple application, we include age and partner age as covariates in the response submodels and additionally include attempt number in both the number of oocytes and double embryo transfer models, and method of fertilization (mixing in vitro or injection with sperm) in the embryo quality submodels. These take on the role of instrumental variables (section 7.4.4) and are included primarily to assist with identification of the model. One change we make to the response model compared to Journal Article 5 is to fit live birth event in the subset of cycles where one or more embryos implanted in the uterine wall, rather than in the subset containing all cycles were embryo transfer was performed. This equates to making a distinction between failure of the transfer due to the embryo not implanting and due to the foetus not being carried to term. This distinction might be important for some research questions (such as the one we tackle in Journal Article 6, Chapter 11). We then define a dropout model for stages t = 1,2,3, corresponding to successful ovarian stimulation, successful fertilization and embryo development, and successful implantation of one or more embryos to the uterus. Accordingly, expanding the model to include a dropout component requires us to add three additional latent probit responses. We fit the model conditional on successful completion of the previous stages, so that only those completing each stage enter into the risk set for the next. We allow the residuals from the dropout model to be correlated with the latent variables in the response model. For the purpose of this illustration, we do not include any covariates in the dropout model. The dataset we use contains 2973 IVF cycles in 2461 women from 2013 to 2015. 12,958 embryos are included. We ignore the clustering of repeated cycles within women here. We develop and fit the model in RStan (Stan Development Team, 2014).

8.4.1. Results of fitting the sequential probit model

Adding the dropout submodel to the fresh cycle model results in slow mixing of the chains used to fit the model (that is, the sampler moves slowly around the posterior distribution). This leads to small effective sample sizes and poor convergence diagnostics for some parameters after running three chains for 3000 iterations each, and discarding the first half as burn in. For example, the regression parameters in the embryo quality submodels corresponding to the upstream response variables ‘number of oocytes’ and ‘fertilisation rate’ have effective sample sizes of around 10 (out of a possible 4500) and Gelman-Rubin convergence statistics around 1.2 (where values below 1.1 might be interpreted as indicating satisfactory convergence). Convergence of the intercepts in the submodels (which include no covariates) is better; effective samples sizes (Gelman-Rubin statistics) for the stimulation, fertilization, and implantation submodels are 794 (1), 1313 (1), and 43 (1.06). The coefficient values correspond to probabilities of 0.9 for progressing beyond stimulation, 0.85 for having sufficient embryos to advance to the transfer stage (given the patient has advanced beyond the stimulation stage) and 0.43 for implantation (given embryo transfer occurs). Dependency between the dropout process and the response model is accommodated through the underlying multivariate latent Normal structure. Examining several of the latent correlation coefficients highlights the difficulties in interpreting the model however. While some of the estimates are coherent (a latent correlation of 0.06 between the number of oocytes obtained and the likelihood of proceeding beyond stimulation), others are less so. For example, we obtain a latent correlation between successful stimulation and achieving embryo transfer (conditional on successful stimulation) of 0.98, with good convergence diagnostics; effective sample size and Gelman-Rubin statistic of 448 and 1.01. It is unclear what this might mean, since the quantity relating to embryo transfer is undefined if stimulation fails. Estimates of the latent correlation between live birth conditional on implantation and both stimulation success and achieving transfer are problematic for similar reasons, although these are based on much smaller effective sample sizes. These observations leave us doubtful as to the utility of jointly modelling a sequential probit model with our fresh cycle model.

8.5 Allowing greater flexibility in the latent variable distribution

All of the MNAR approaches presented here, as well as the MAR approach (that is, our 'do nothing' approach, where we consider our response model to be sufficient without any additional dropout component) involve estimation of a latent variable structure tying observed to unobserved responses in those who dropped out. Whether we allow the relationship to depend on dropout (MNAR) or not (MAR), it might be preferable not to estimate latent variables corresponding to unrealised responses, nor the corresponding relationships with upstream response variables which actually preclude them. To this end, we might consider approaches where the latent variable distributions are allowed to vary with drop out, so that patients who do not undergo transfer do not have transfer outcomes included in their latent variable vector. The relationships between observed stages could then vary according to the stage of dropout. This would bear some resemblance to the pattern-mixture models described by Little (1993), or, more recently, approaches using mixtures of latent variables to allow for different subgroups of patients (Komarek, et al., 2010). We return to this in the discussion to the thesis.

8.6 Summary of Chapter 8.

In this chapter, we discussed the relatively subtle matter of incorporating drop out into our models of the fresh IVF cycle. We highlighted the likelihood that the appropriate target of missing data modelling is likely to be context-dependent, as well as the fact that it might be appropriate to adopt different approaches for treating drop out at different stages of the cycle. While we reviewed several methods that would allow missing data due to drop out to be MNAR, we queried the suitability of these analyses, on several grounds. First, drop out from the cycle is usually due to poor response or outright failure at one of the stages. These responses are included in our model, such that drop out is likely to be ignorable given the observed data (MAR). While measurement error and model misspecification weakens our ability to assert MAR, under this view the correct solution lies in identifying and using good quality measurements and flexible representations of them wherever they appear as outcome-covariates (for example, by using splines), rather than by using special analysis methods. Another conceptual difficulty is the fact that, due to the sequential ordering of the responses in the model, once a patient drops out from the cycle, the responses at remaining stages are not defined. The patient's multivariate response vector is therefore truncated rather than censored. Under this view, the MNAR methods described here are not appropriate, since they assume some underlying relationship between responses pre and post-dropout. This was highlighted in an application of the sequential probit model, which yielded correlation coefficients that were not obviously interpretable. Approaches exist which allow for greater flexibility in the latent variable distribution. It may be possible to adapt these to the present setting, so as to alleviate some of this conceptual baggage. This remains a topic for future research.

8.7 **References for Chapter 8.**

Alviggi C, Cognigni GE, Morgante G, Cometti B, Ranieri A, Strina I, Filicori M, De Leo V, De Placido G. A prospective, randomised, investigator-blind, controlled, clinical study on the clinical efficacy and tolerability of two highly purified hMG preparations administered subcutaneously in women undergoing IVF. *Gynecol Endocrinol* 2013;29: 695-699.

Check JH, Bollendorf A, Summers-Chase D, Yuan W, Horwath D. Isolating sperm by selecting those with normal nuclear morphology prior to intracytoplasmic sperm injection (ICSI) does not provide better pregnancy rates compared to conventional ICSI in women with repeated conception failure with in vitro fertilization. *Clin Exp Obstet Gyn* 2013;40: 15-17.

Diggle P. Analysis of longitudinal data. 2nd edn, 2002. Oxford University Press, Oxford ; New York.

Diggle P, Farewell D, Henderson R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *J R Stat Soc Ser C Appl Stat* 2007;56: 499-529.

Diggle P, Kenward MG. Informative Drop-out in Longitudinal Data-Analysis. *J R Stat Soc Ser C Appl Stat* 1994;43: 49-93.

Ferraretti AP, Gianaroli L, Motrenko T, Feliciani E, Tabanelli C, Magli MC. LH pretreatment as a novel strategy for poor responders. *BioMed Research International* 2014;2014: 926172.

Friedler S, Ben-Ami I, Gidoni Y, Strassburger D, Kasterstein E, Maslansky B, Komarovsy D, Bern O, Ron-El R, Raziel A. Effect of seminal plasma application to the vaginal vault in in vitro fertilization or intracytoplasmic sperm injection treatment cycles-a double-blind, placebo-controlled, randomized study. *J Assist Reprod Genet* 2013;30: 907-911.

Gao X, Chang XF, Du HL, Zhang M, Zhang JP, Zhu AP. Effect of soothing liver therapy on oocyte quality and growth differentiation factor-9 in patients undergoing in vitro fertilization and embryo transfer. *J Tradit Chin Med* 2013;33: 597-602.

Hogan JW, Scharfstein DO. Estimating causal effects from multiple cycle data in studies of in vitro fertilization. *Stat Methods Med Res* 2006;15: 195-209.

Komarek A, Hansen BE, Kuiper EM, van Buuren HR, Lesaffre E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Stat Med* 2010;29: 3267-3283.

Little RJA. Pattern-Mixture Models for Multivariate Incomplete Data. *J Am Stat Assoc* 1993;88: 125-134.

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.

McConnell S, Stuart EA, Devaney B. The truncation-by-death problem - What to do in an experimental evaluation when the outcome is not always defined. *Evaluation Rev* 2008;32: 157-186.

Revelli A, Chiado A, Dalmasso P, Stabile V, Evangelista F, Basso G, Benedetto C. "Mild" vs. "long" protocol for controlled ovarian hyperstimulation in patients with expected poor ovarian responsiveness undergoing in vitro fertilization (IVF): a large prospective randomized trial. *J Assist Reprod Genet* 2014;31: 809-815.

Rubin DB. Inference and Missing Data. *Biometrika* 1976;63: 581-590.

Rubin DB. Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death*.* *Stat Sci* 2006;21: 299-309.

Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models, 2004. Crc Press.

Soullier N, Bouyer J, Pouly JL, Guibert J, de La Rochebrochard E. Estimating the success of an in vitro fertilization programme using multiple imputation. *Hum Reprod* 2008;23: 187-192.

Stan Development Team. RStan: the R interface to Stan, Version 2.5.0. 2014.

Steele F, Sigle-Rushton W, Kravdal O. Consequences of Family Disruption on Children's Educational Outcomes in Norway. *Demography* 2009;46: 553-574.

Steele F, Washbrook E. Discrete-time Event History Analysis. Centre for Multilevel Modelling University of Bristol. 2013.

Verberg MF, Eijkemans MJ, Heijnen EM, Broekmans FJ, de Klerk C, Fauser BC, Macklon NS. Why do couples drop-out from IVF treatment? A prospective cohort study. *Hum Reprod* 2008;23: 2050-2055.

Waelbroeck P. Computational issues in the sequential probit model: a Monte Carlo study. *Computational Economics* 2005;26: 141-161.

Zhang JNL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J Educ Behav Stat* 2003;28: 353-368.

Chapter 9. What are the sources of variation in controlled ovarian stimulation and what are the implications for personalisation of ART? A multilevel modelling study

Journal article 4

Authors Oybek Rustamov↑, Jack Wilkinson↑, Antonio La Marca, Cheryl Fitzgerald, Stephen A. Roberts

↑ The contributions of these two authors was equal, and they should be considered joint first authors.

Status Under peer review at Human Reproduction Open

Contribution statement OR prepared the study protocol, prepared the dataset, conducted and interpreted preliminary statistical analysis, and coauthored the manuscript. JW conducted and interpreted statistical analysis and coauthored the manuscript. SR and CF oversaw and supervised preparation of datasets, statistical analysis, contributed to the interpretation of statistical analysis and coauthored the manuscript. ALM contributed to the interpretation of statistical analysis and coauthored the manuscript.

Preamble In this article, we use the stimulation model developed in Chapter 6. to investigate the scope for personalisation of ovarian response. We use multilevel models to quantify the amount of known and unknown variation in stimulation responses, as well as the proportion of variation attributable to modifiable treatment factors. The study complements recent RCTs of personalised stimulation algorithms, by illustrating the limited scope for prediction and manipulation of ovarian response on the basis of current knowledge.

Outputs and Impact of the research This work resulted in a collaboration with Prof. Antonio La Marca of the Mother-Infant Department, University of Modena and Reggio

Emilia, Modena, Italy. ALM is a leading expert in the area of personalized ovarian stimulation. In addition to the present study, JW and ALM are now collaborating on a Cochrane Review on this topic.

9.1 Abstract

Study question

How much variation in COS response can be accounted for by known patient and treatment characteristics, and what are the implications for individualised stimulation protocols?

Summary answer

There is substantial variation in the COS responses of similar women and in repeated COS episodes undertaken by the same woman, which cannot be accounted for at present. This suggests that there is likely to be limited scope for personalised treatment unless additional predictors of ovarian response can be identified.

What is known already

The goal of individualized COS is to safely collect enough oocytes to maximise the chance of success in an ART cycle. Personalisation of treatment rests on the ability to reduce variation in response through modifiable factors.

Study design, size, duration

Multilevel modelling of a routine ART database covering the period 1st October 2008 to 8th August 2012 was employed to estimate the amount of variation in COS response and the extent to which this could be explained by immutable patient characteristics and by manipulable treatment variables. 1851 treatment cycles undertaken by 1430 patients were included. The study was not subject to attrition, as cancelled cycles were included in the analysis.

Participants/materials, setting, methods

Women of 21-43 years of age undergoing ovarian stimulation for IVF (possibly with ICSI) using their own eggs at the Reproductive Medicine Department of St Mary's Hospital, Manchester, England.

Main results and the role of chance

Substantial unexplained variation in COS response was observed (3.4-fold (95% CI: 3.12 to 3.61)). Only a relatively small amount of this variation (around 19%) can be explained by modifiable factors. A significant, previously undescribed predictor of response was the practitioner performing oocyte pickup, with 1.5 fold variation between surgeons with the highest and lowest yields.

Limitations, reasons for caution

Although a large number of covariables were adjusted for in the analysis, including those that were used for dosing and determination of the stimulation regimen, this study is subject to confounding due to unmeasured variables and measurement error.

Wider implications of the findings

The present study suggests that there are limits to the extent that COS response can be predicted on the basis of known factors, or controlled by manipulation of treatment factors. Moreover, modifiable variation in response appears to be partially attributable to differences between surgeons performing oocyte pick up. Consequently, consistent prevention of ineffective or unsafe responses is not likely to be possible at present. Our results highlight the importance of blinding surgeons in RCTs.

KEYWORDS

Ovarian Stimulation, Assisted Reproductive Technology, Ovarian Response, In Vitro Fertilisation.

9.2 Introduction

The goal of controlled ovarian stimulation (COS) in ART is to safely obtain enough oocytes to maximize the chance of success in the treatment cycle. Frequently, this goal proves elusive; it has been estimated that 17% of ART stimulation cycles in the UK (Sunkara, et al., 2011) and 28% in the US (Steward, et al., 2014) result in the collection of over 15 oocytes, representing increased risk to both the woman (Steward, et al., 2014) and any potential offspring (Sunkara, et al., 2015). In total, around 12% of IVF cycles in the UK are cancelled due to poor or excessive ovarian response (Kurinczuk, 2010). If this situation is to be improved, methods to predict and prevent ineffective or unsafe COS responses are required (La Marca, et al., 2012, La Marca and Sunkara, 2014). To this end, the predictive value of two ovarian reserve tests (ORT), anti-mullerian hormone (AMH) and antral follicle count (AFC), has been demonstrated in relation to COS response (Broer, et al., 2011, Broer, et al., 2013). In addition, the dose-responsiveness of COS response to follicle-stimulating hormone (FSH) has also been established (Arce, et al., 2014), although this is likely to be limited to patients with sufficient ovarian reserve to permit tailoring (Klinkert, et al., 2005, Lekamge, et al., 2008). The value of ovarian reserve testing for improving clinical outcomes of ART is less clear however, with a recent review of RCTs of individualized versus standard doses of FSH noting that only one trial in good prognosis patients had demonstrated an effect on pregnancy (van Tilborg, et al., 2016). The same review concluded that tailoring the dose of FSH on the basis of ORTs may improve safety, however. Some support for this is provided by a recent RCT where a multivariable dose selection algorithm increased the proportion of participants obtaining an optimal number of oocytes, albeit using a definition that was not prespecified (Allegra, et al., 2017). A second RCT suggested that dose-selection using AMH may reduce the overall proportion of low or excessive responses, although these analyses excluded cycles cancelled for poor response (which occurred more frequently in the personalized group) (Nyboe Andersen, et al., 2017).

From a statistical perspective, we contend that the challenge of optimizing COS should be viewed as the need to reduce variation in response. This is somewhat different to the typical situation we face when designing and testing interventions, where effectiveness is defined as a shift in an outcome in one direction. In this regard, an understanding of the sources of variation contributing to the distribution of COS outcomes would be advantageous (Senn, 2016). In particular, the amount of unexplained variation represents a limit on our ability to predict response under a given treatment regimen, and the degree to which we can manipulate this response depends on the amount of variation attributable to modifiable factors. This in turn motivates the identification of additional sources of heterogeneity which may be incorporated into multivariable prediction and tailoring algorithms. Moreover, quantifying the degree of variation associated with known predictors highlights variables to be controlled in clinical practice and in research. While RCTs should, in principle, produce balance over nuisance factors between treatment arms, in reality the impracticability of blinding these trials undermines this in the form of performance biases (Higgins, et al., 2011).

Multilevel modelling is a statistical technique that allows us to attribute variation to known and unknown factors, whilst estimating and allowing for measured covariate effects. The variation of unknown source can be apportioned to 'between-patient' (factors that are intrinsic to the patient) and 'within-patient' (factors which might vary between repeated treatment cycles) components (Snijders and Bosker, 2012). In order to investigate the impact of known and unknown sources of variation on COS response, we constructed multilevel models using a large routine ART database. We discuss the implications for practice and research of individualised COS.

9.3 Materials and methods

9.3.1. Population

Women of 21-43 years of age undergoing ovarian stimulation for IVF (possibly with ICSI) using their own eggs at the Reproductive Medicine Department of St Mary's Hospital, Manchester from 1st October 2008 to 8th August 2012 were included. Patients that had AMH measured using only the Gen II assay were excluded, given previously reported problems with this assay (Rustamov, et al., 2012). Patients with ultrasound features of polycystic ovaries, previous history of salpingectomy, ovarian cystectomy and/or unilateral salpingoophorectomy were excluded from the analysis as we expected the relationships between patient and treatment characteristics and response to be distinct in these subgroups. Similarly, small numbers of cycles with ovarian stimulation other than GnRH agonist long down regulation or Short GnRH antagonist cycles were not included in the study.

Patients with a history of unilateral tubal occlusion or unilateral salpingectomy were categorized as mild tubal factor infertility and patients with blocked tubes bilaterally or with history of bilateral salpingectomy were classified as having severe tubal disease. Severe male factor infertility was defined as the partner having azoospermia, surgical sperm extraction or severe oligospermia, which necessitated Multiple Ejaculation Resuspension and Centrifugation test (MERC) for assisted conception. Mild male factor was defined as abnormal sperm count that did not meet the aforementioned criteria for severe male infertility. Diagnosis of endometriosis was based on a previous history of endometriosis confirmed using Laparoscopy. Diagnosis of endometrioma was established using a transvaginal ultrasound scan prior to IVF treatment. In couples without a definite cause for infertility following investigation, the diagnosis was categorized as unexplained.

9.3.2. **Measurement of AMH and AFC**

AMH measurements were performed by the Clinical Assay Laboratory of Central Manchester NHS Foundation Trust, and the procedure for sample handling and analysis was based on the manufacturer's recommendations. Venous blood samples were taken without regard to the day of women's menstrual cycle and serum samples were separated within two hours of venipuncture. Samples were frozen at -20C until analysed in batches using the enzymatically amplified two-site immunoassay (DSL, Active MIS/AMH ELISA; Diagnostic Systems Laboratories, Webster, Texas). The intra-assay coefficient of variation (CV) (n=16) was 3.9% (at 10pmol/l) and 2.9% (at 56pmol/l). The inter-assay CV (n=60) was 4.7% (at 10pmol/l) and 4.9% (at 56pmol/l). Haemolysed samples were not included in the study. In patients with multiple AMH measurements, the value closest to their IVF treatment cycle was selected. The working range of the assay was up to 100pmol/L and a minimum detection limit was 0.63pmol/L. Results falling below the minimum detection limit were coded as 50% of the minimum detection limit (0.31 pmol/L) and test results that were higher than the assay ranges were coded as 150% of the maximum range (150 pmol/L).

In our department, the measurement of AFC is conducted as part of an initial clinical investigation before the first consultation with clinicians and prior to the IVF cycle. Qualified radiographers performed the assessment of AFC during the early follicular phase (Day 0-5) of the menstrual cycle. Measurement of AFC consisted of the counting of all antral follicles measuring 2-6mm in longitudinal and transverse cross sections of both ovaries using a transvaginal ultrasound scan. The AFC measurement closest to the date of the IVF cycle was selected for the analysis.

9.3.3. Description of COS Protocols

On the basis of their AMH measurement, patients were stratified into the treatment bands for ovarian stimulation using COS protocols. During the study two different COS protocols were used and in addition three minor modifications were made in the 2nd protocol. Time periods, AMH bands, down regulation regimes, initial dose of gonadotropins and adjustment of daily dose of gonadotropins for each protocol are described in S Table 18. Similarly the management of excessive ovarian response was tailored to pretreatment AMH measurements, although mainly based on the results of oestradiol and scan monitoring during the stimulation period (S Table 18). Assessment of transvaginal ultrasound guided follicle tracking and serum oestradiol levels on specific days of the stimulation were used for monitoring of COS (S Table 19). The criteria for the cycle cancellation for poor ovarian response were consistent across all protocols; fewer than 3 follicles >15mm in size on Day 10 of ovarian stimulation.

9.3.4. Pituitary desensitisation regimes

Selection of pituitary desensitisation regime was based on the patient's AMH according to the COH protocol at the time of commencement of the IVF cycle (S Table 18). Long agonist regimes involved daily subcutaneous injection of 250μg or 500 μg of the GnRH agonist Buserelin acetate (Supercur, Sanofi Aventis Ltd., Surrey, UK) from the mid-luteal phase (Day 21) of the preceding menstrual cycle, which continued throughout ovarian stimulation. Women treated with Antagonist regime had daily subcutaneous administration of GnRH antagonist Ganirelex (Orgalutran, Organon Laboratories Ltd., Cambridge, UK) from Day 4 post-stimulation until the day of HCG/GnRH agonist trigger. Ovarian stimulation was achieved by injection of daily dose of hMG, Menopur (Ferring Pharmaceuticals, UK) or rFSH, Gonal F (Merck Serono) as per the AMH-tailored protocols (S Table 18). Oocyte maturation was triggered using 5000 international units of HCG

(Pregnyl, Organon Laboratories Ltd., Cambridge, UK) and the criteria for timing of HCG injection was consistent across all protocols: one (or more) leading follicles measuring >18mm and two (or more) follicles >17mm.

9.3.5. Oocyte collection

Oocyte collection was conducted 34-36 hours following injection of HCG for follicle maturation. An ultrasound guided oocyte pick up (OPU) was conducted by experienced clinicians under sedation. Practitioners with a small number (<10) of oocyte collection procedures were pooled in the analysis (group J). If the cycle was cancelled before oocyte recovery, it was categorized under the practitioner who was on-call for oocyte recovery on the day of cycle cancellation.

Oocytes were counted immediately post-OPU by an embryologist. In patients undergoing ICSI, the assessment of the quality of oocytes was conducted 4-6 hours post-OPU. Oocytes assessed as in Metaphase II stage (MII) of maturation were categorized as mature.

9.3.6. Study outcomes

We evaluated the outcome number of oocytes recovered and, in the subset of patients undergoing ICSI, the number of mature oocytes. However, our estimates relating to mature oocytes were so imprecise as to be quite uninformative. Consequently, we present these without further comment.

9.3.7. Statistical analysis

We used multilevel multivariable Poisson regression to estimate the effects of patient and treatment characteristics on stimulation response (Snijders and Bosker, 2012). The variables included in the regression models were selected on the basis of background knowledge and the objectives of the study. We distinguished patient characteristics (age, AMH, AFC, BMI, attempt number and cause of infertility) which cannot be altered from treatment variables (initial dose of gonadotropin, stimulation regime (antagonist or long agonist), protocol (old version or v1, v2 & v3 or v4 of the new protocol), type of gonadotropin (HMG or rFSH) and OPU practitioner, which could in principle be used to tailor treatment. The representation of age, AMH and AFC in the model was determined on the basis of exploratory analysis consisting of graphing each variable against egg count and log(egg count), and by comparing models featuring competing representations using Akaike's Information Criterion (Akaike, 1972), a measure of fit that penalizes complexity. As a result of this process, age was represented as a quadratic in the final analysis, AMH was log-transformed and AFC was categorized into 3 levels on the basis of quantiles. Initial dose of gonadotropin was represented as a categorical variable; this decision was made on the basis of the distribution of the doses and the desire to obtain an easily interpretable model (Table 13). Interactions between regime and other variables and dose and other variables were considered using likelihood ratio testing and graphing of the predictors against egg count within regime and dose categories. Dose effect was allowed to vary with regime in the final analysis, owing to the observed significance of this interaction using a likelihood ratio test and the inherent plausibility of this relationship. We also fitted a version of the final model with an interaction between log(AMH) and dose, to investigate whether the relationship between dose and oocyte yield varied with AMH level. Continuous variables were mean-centered and standardized by dividing by a standard deviation. This was done for the purposes of interpretability and to improve computational efficiency in model fitting.

Poisson regression models for oocyte yield and number of mature oocytes (for ICSI cycles only) as outcome variables were fitted for the final analysis with multiplicative random effects at both the observation and patient-levels included to account for the high variability in cycle outcomes and the correlation between repeated cycles undertaken by the same patient, respectively. This method produces covariate-adjusted yield ratios and 95% CIs. For categorical variables, these can be interpreted as relative yields per cycle for each level of the predictor compared to a reference category. For continuous variables, they can be interpreted as the multiplicative change in the yield per cycle associated with a standard deviation increase in the predictor. We used multiple imputation to handle the relatively low proportion of missing values in the dataset (see Table 13), including imputed egg counts for cancelled cycles. All of the variables included in the analysis were used in the imputation process, in addition to variables relating to follicle counts on days 8 and 10 of the stimulation phase and the total dose of gonadotropins administered. We examined plots of residuals and of predictions arising from the analysis to assess model fit. Analysis was conducted using the software packages R (R Core Team, 2014) and RStan (Stan Development Team, 2014). Imputation was conducted using the mi package (Su, et al., 2011). No sample size calculation was performed, as we were not interested in hypothesis testing. Instead, we rely on 95% CIs to indicate the precision of our results. We estimated the amount of unexplained between and within-patient variation, and of total variation, in three models of oocyte yield: 1) no covariates; 2) patient covariates only; and 3) treatment and patient covariates. The first of these quantifies the variance in the data. By comparing model 1 to model 2 we can estimate the amount of variation attributable to patient characteristics and by comparing model 2 to model 3 we estimate the amount that could, in principle, be reduced through treatment. We used the distribution of the random effects from the fitted models to calculate these measures of unexplained variation. Each model yields two random effects for each patient in the analysis, which describe how each patient's responses differ relative to the outcome that would be expected according to the model variables (patient and cycle-specific yield ratios). We calculated the yield ratio for a random effect one standard deviation above the mean (YR_{SD}), the ratio of the 95th to the 5th random effects (YR_{90}), and the variance of the random effects for each model, overall and partitioned as within and between patients.

9.4 Results

9.4.1. Characteristics of the sample

The dataset contained 1851 treatment cycles (defined as initiation of COS) on 1430 patients. 1070 (75%) patients had one cycle, 306 (21%) had two, 56 (4%) had three and 1 (0%) had four. 1236 ICSI cycles on 964 patients were available for the analysis of mature oocytes. Table 13 gives a summary of the characteristics of the cycles in the dataset.

9.4.2. How much variation in COS response is explained by immutable patient characteristics?

Table 14 shows measures of unexplained variation (YR_{SD}, YR_{90}, and the residual variance, see *Statistical Analysis*) in three models of COS response.

The reduction in these measures between models 1 and 2 tells us how much is explained by patient characteristics. It is evident that patient characteristics explain a substantial portion of the overall variation; the total unexplained variance (the sum of the between and within-patient components) reduces from 0.30 to 0.16 (that is, to 53% of the original value) when these are added. This translates to a YR_{SD} of 1.75 in model 1 compared to 1.51 in model 2. The YR_{90} is 6.30 in model 1 and 3.89 in model 2. We can see that known patient characteristics explain variation through the between-patient rather than the within-patient component (as there is no substantive reduction in the latter, Table 14). This is unsurprising, since these variables tend not to vary from cycle to cycle.

Characteristic	Summary
Total dose of gonadotropins (IU)	3000 2100 to 3300 300 to 7650 *0%*
Initial dose of gonadotropins (IU)	*0%*
75-150IU	297 (16)
187-250IU	484 (26)
300IU	919 (50)
375IU	62 (3)
450IU	89 (5)
Age at start of cycle (years)	33.7 30.3 to 36.9 21.5 to 43.7 *0%*
BMI at start of cycle	24.0 21.5 to 26.8 16.3 to 36.0 *15%*
AMH at start of cycle (pmol/L)	15.0 9.4 to 22.7 1.3 to 150 *0%*
Regime	*0%*
Long Agonist	821 (44)
Antagonist	1030 (56)
Gonadotrophin	*0%*
HMG	1602
rFSH	233
AFC	13 10 to 17 3 to 52 *10%*
Attempt no	*0%*
1	1347 (73)
2	409 (22)
3	91 (5)
4	4 (0)
Number of eggs recovered (cancelled cycles set to missing)	9 5 to 14 0 to 38 *2%*

Table 13: Summary of cycle characteristics. Median, IQR and range for continuous variables, frequency and percentage for categorical variables. % missing shown in italics.

9.4.3. How much variation in COS response can be explained by manipulable treatment factors?

Similarly, a comparison between models 2 and 3 shows how much variation can be accounted for by treatment (Table 14). Adding treatment variables to the model does reduce overall variation further, but only modestly. Total variance reduces from 0.16 to 0.13 (81% of the original). The YR_{SD} are 1.51 and 1.45 in the models 2 and 3 respectively, and the YR90 are 3.87 and 3.36. As such, the model implies that there is a limit to the extent to which variation in response can be reduced by tailoring treatment, with the YR90 of 3.4 implying that a greater than three-fold difference in yield could reasonably be observed between two cycles in which two patients with similar characteristics are treated in the same way. If the same patient were to be treated in the same way on two occasions, a 2.7-fold difference in yield could reasonably be observed. This can be translated to a clinically meaningful scale. Suppose that a patient obtained 9 eggs from a cycle. If another patient with similar characteristics were to be treated in the same way, we would expect their response to be between 6 and 13 eggs (based on YR_{SD}), although any response in the range 4 to 19 (based on YR_{2SD}) would not be surprising. If the same initial patient were stimulated in the same way a second time we would expect a response between 7 and 12 eggs, but any response between 5 and 17 eggs should be anticipated.

9.4.4. Effects of known patient and treatment characteristics

Yield ratios with 95% CIs from the fitted models are presented visually in Figure 35 and in S Table 20. The corresponding estimates for the analysis of mature oocytes are displayed in S Figure 1 and S Table 20. These refer to the estimated 'effects' of the predictor variables on COS response, as described in *Statistical Analysis*, above. Notably, the ratio of the greatest to the lowest yield ratio estimated for the practitioners was 1.53, with differences between operators apparent on the basis of non-overlapping 95% CIs (Figure 35). Whilst AMH was a strong predictor of response, we did not find evidence of differential effects of AMH across

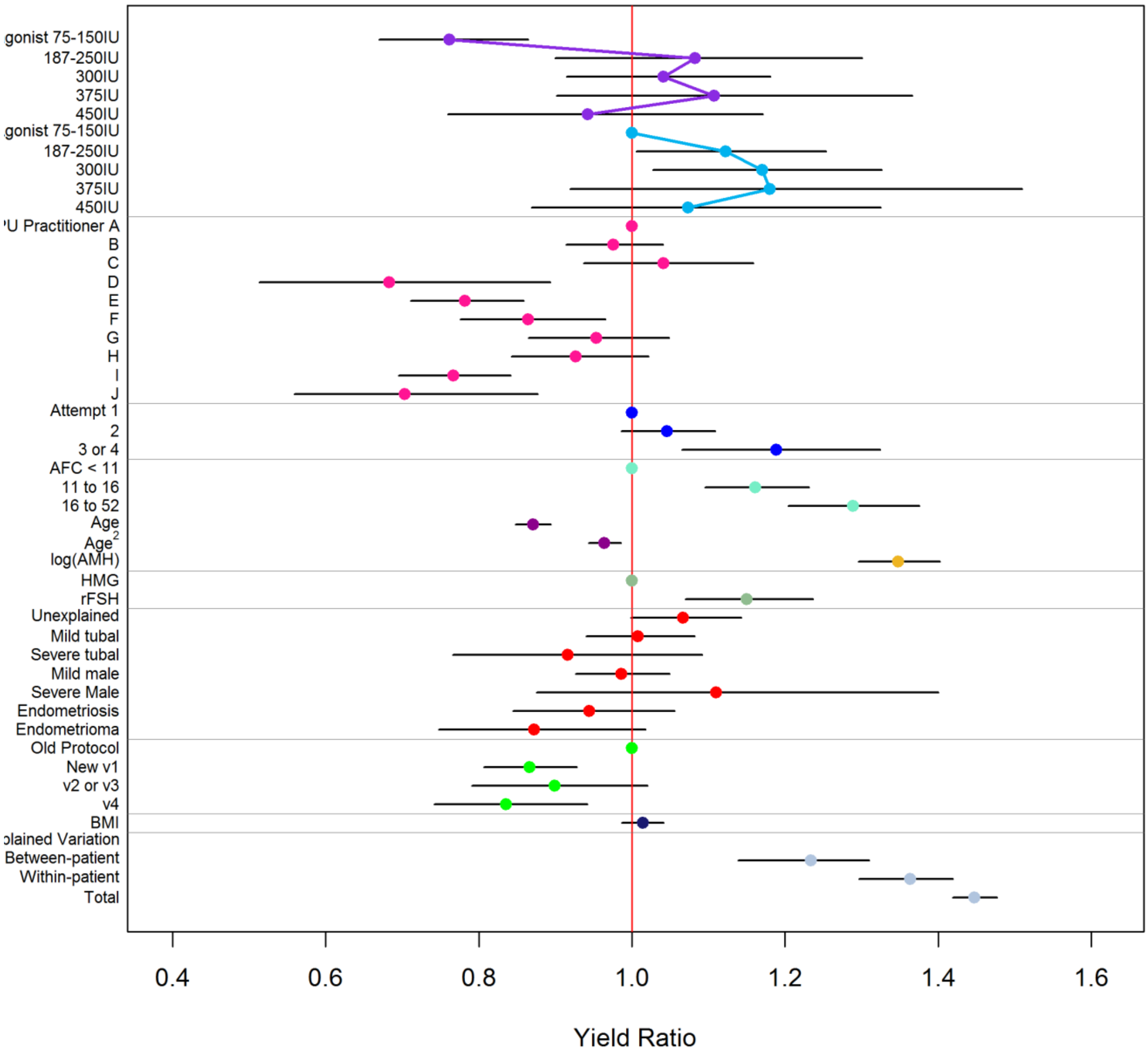


Figure 35: Yield ratios and 95% CIs from the multivariable Poisson regression model of number of oocytes per cycle. Continuous predictors have been standardized, so that coefficients display the expected multiplicative increase in the yield ratio for a standard deviation change in the variable. Increasing dose effect under a GnRH Antagonist regime is shown by the purple connecting line. Increasing dose effect under a GnRH Long Agonist regime is shown by the blue connecting line.

dose groups (Interaction test: p = 0.60), although our power to detect such an effect is likely to have been low. Other predictor variables showed effects in the anticipated directions, with increased yields for higher AFC values and decreased yields for increasing age, for example. The model suggested increased yields when rFSH was used compared to an equivalent starting dose of HMG.

Model	Random effect YR for +1 SD vs mean (YR_{SD})			Random effect variance			Ratio of 95th to 5th quantile of random effect YRs (YR_{90})		
	Between-patient	Within-patient	Total	Between-patient	Within-patient	Total	Between-patient	Within-patient	Total
1: No covariates	1.55 (1.45 to 1.63)	1.43 (1.36 to 1.52)	1.75 (1.72 to 1.80)	0.18 (0.13 to 0.22)	0.12 (0.09 to 0.16)	0.30 (0.27 to 0.33)	4.15 (3.35 to 4.90)	3.29 (2.76 to 3.98)	6.30 (5.84 to 6.83)
2: Patient covariates w attempt	1.19 (1.08 to 1.28)	1.45 (1.39 to 1.51)	1.51 (1.48 to 1.54)	0.03 (0.01 to 0.06)	0.13 (0.10 to 0.16)	0.16 (0.14 to 0.18)	1.78 (1.32 to 2.25)	3.39 (2.89 to 3.89)	3.87 (3.61 to 4.15)
3: Patient plus treatment covariates	1.23 (1.14 to 1.31)	1.36 (1.30 to 1.42)	1.45 (1.42 to 1.48)	0.04 (0.02 to 0.07)	0.10 (0.07 to 0.12)	0.13 (0.12 to 0.15)	1.98 (1.53 to 2.40)	2.70 (2.31 to 3.16)	3.36 (3.12 to 3.61)

Table 14: Measures of unexplained variation (95% CIs) in three models of oocyte yield.

9.5 Discussion

In the present study, we used multilevel modelling of a routine ART database to quantify the various sources of variation in response to COS. Our results quantify, and are consistent with, the effects of known predictors (Figure 35, S Table 20), while large random effects (yield ratios) suggest that there remains substantial variation that we cannot currently account for (a 3.4 fold difference Figure 35, Table 14). This holds both for differences between the responses of different women and between repeated responses of the same woman. Only a relatively small amount of this variation (around 19%) can be explained by modifiable treatment factors.

9.5.1. Patient characteristics

Patient characteristics explained a substantial portion of variation between women. This included strong relationships with known measures of ovarian competence (age, AMH, AFC)(La Marca and Sunkara, 2014). Variation in BMI was quite precisely estimated as having little to no influence on oocyte yield, possibly because all patients had values in the range 19 to 30. There was no evidence to suggest that any particular infertility diagnosis was associated with number of oocytes, with the exception of increased yields (estimate of 7%, no higher than 14%) for those with unexplained infertility. Number of oocytes appeared to increase with attempt number, with increased yields for 2nd attempts and subsequent attempts. This could reasonably be an artefact due to selection effects relating to different profiles or treatment strategies for patients undergoing multiple treatment attempts, although a sensitivity analysis excluding attempt number had no discernible impact on the other model estimates or on the amount of explained variance.

9.5.2. Treatment characteristics

This appears to be the first study to identify a substantial effect of oocyte recovery practitioner on oocyte yield. It is worth noting that the operators were all trained, experienced surgeons. Whilst tailoring of the allocation of patients to practitioner lacks credibility as a treatment protocol, this variability does suggest that there are as yet unmeasured factors which affect COS outcome which if identified may have the potential for optimisation. This finding is important, as variation due to recovery practitioner could undermine any attempts to guide a patient to an optimal oocyte yield by tailoring the gonadotropin dose. Blinding of the recovery practitioner and recording of the allocation of patients to practitioner should be a mandatory feature of RCTs of personalized COS.

In line with previous research in this area (Arce, et al., 2014), the model suggested a dose-response relationship between initial gonadotropin dose and number of oocytes at lower doses. However, this did not appear to be sustained beyond the lowest dose. This suggests that, to the extent that tailoring the dose is possible, it should be restricted to a lower dose range (Figure 35). Differences between antagonist and long agonist regimens were generally unclear, other than for the 75-150 IU dose band where we observed a reduced number of oocytes in antagonist cycles. In order to translate dose and regimen effects to a more easily interpretable scale, we plotted the observed oocyte yields together with the predicted oocyte yields from our model for patients falling in low, medium and high AMH bands, using cut-offs of <5pmol/L, 5-15pmol/L and >15pmol/L, which have been suggested (Nelson, et al., 2007) and used (Nelson, et al., 2009) elsewhere in the literature (Figure 36). This represents the predicted outcomes for our centre, were dose selection performed solely on the basis of AMH. This figure highlights the impact of other sources of variation that should be considered in individualised COS, because the variation within each AMH/protocol/dose category is large relative to the variation between categories , and suggests that multivariable algorithms (La Marca, et al., 2012, Popovic-Todorovic, et al., 2003) will be needed to obtain reliable predictions of response. However, our models also suggest that many of these contributive variables remain unknown. We did not replicate the finding of Arce and colleagues (2014) that dose effects vary according to AMH, although our power to detect an effect of this nature is likely to have been low. The predictions appear to be consistent with existing research and writing on this topic, indicating in particular that increasing the dose in patients with predicted low response is unlikely to increase the oocyte yield (Klinkert, et al., 2005, Lekamge, et al., 2008) and that dose-effects on the mean response are modest (Sterrenburg, et al., 2011).

In this case, the effect on the mean response may not represent the most useful measure of efficacy however. Given that the goal of individualised COS is to prevent insufficient or unsafe responses (La Marca et al., 2012), we believe that it is most useful to focus on the effects of interventions on reducing variation in outcome. In this context, an intervention could be ‘effective’ even if no effect on the mean was observed. Our analysis suggests that treatment differences account for relatively little of this variation, and this is likely to limit the extent to which extreme responses can be prevented by tailoring treatment. A unidirectional mean effect will of course be more relevant in populations of expected poor or high responders compared to unselected patients, although even then a simple ‘mean difference’ may conceal deleterious consequences of treatment (if, for example, more expected high (low) responders end up having poor (excessive) responses, as appears to be the case in Nyboe Andersen, et al., 2017). As a result, many trials quantify COS response by categorizing responses as ‘poor’, ’normal’ or ‘high’, and use this as a trial endpoint (eg: Allegra, et al., 2017; van Tilborg, et al., 2012; Popovic-Todorovic, et al., 2003). This is not entirely unreasonable if the criteria are predefined and cancelled cycles are included in the denominator, although categorizing measurements in this way reduces power in the trial, necessitating larger sample sizes (Altman and Royston, 2006).

We note that simple statistical methods exist for comparing variation between treatment arms directly, such as Levene's test (Schultz, 1985).

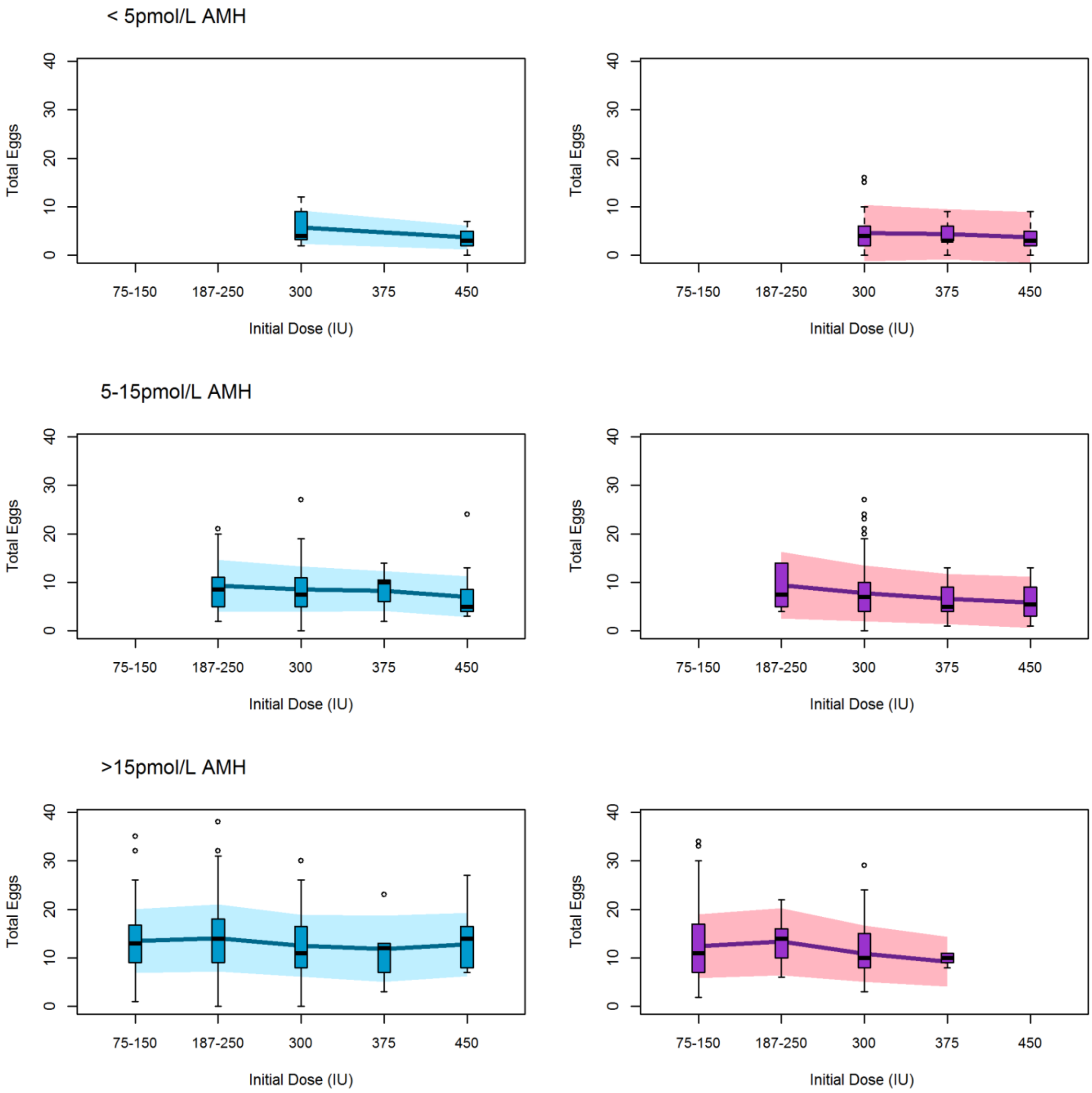


Figure 36: Distribution of observed egg counts (box and whisker plots) with those predicted under the model for low (DSL assay < 5 pmol/L), medium (5-15 pmol/L) and high (>15 pmol/L) AMH bands for both GnRH Long Agonist (blue) and GnRH Antagonist (purple) regimes. Solid line represents the mean response from the posterior predictive distribution. Shaded area represents +/- 1 SD. Note that other covariate values are not fixed but reflect the characteristics of the sample. Only groups with 5 or more observations are shown.

Limitations of the present study should be noted. There may be concerns over the generalizability of our findings, since some of the doses administered in the dataset are higher than would typically be used, for example, throughout Europe. However, we note here that our concern is not in the evaluation of any particular treatment strategy, but rather to tease apart the contributions of various predictors on COS response. Regardless, we conducted a sensitivity analysis where we fitted a model in the subset of participants treated with doses of 225IU or lower (S Figure 2). The estimates are consistent with our main analysis, albeit with reduced precision due to the reduction in sample size. While we included a large number of predictor variables, there is likely to be confounding due to unmeasured predictors as well as 'residual confounding' due to measurement error in the model covariates (Sterne, et al., 2016). In particular, there may be concern around confounding by indication due to selection for treatment on the basis of prognosis (Walker, 1996). In this regard, we note that we have included all of the variables that were used for treatment allocation in the model (at least in principle), and measures of balance between dose groups (McCaffrey, et al., 2013) suggest a reasonable degree of balance after adjusting for covariates, other than for the highest versus the lowest dose band.

We suggest that an understanding of the degree and determinants of variation in COS response is key to improving clinical practice and conducting research in this area. The goal of personalized COS is to reduce this variation, and this may be assisted both by incorporating a range of predictive patient characteristics into dose algorithms and by attempting to standardize aspects of treatment that may introduce noise (Senn, 2016). Our results indicate that much of the variation in response cannot be explained by known factors however. We have identified the oocyte recovery practitioner as one potential source of variation in this study, and recommend that blinding is used in RCTs to reduce associated performance biases. Moreover, we advise that the allocation of participant to practitioner is recorded and considered as a covariate in any analysis. We conclude that, until additional predictors of variation are identified, consistent prevention of extreme responses is unlikely to be achieved.

Funding

JW is funded by a Doctoral Research Fellowship from the National Institute for Health Research (DRF-2014-07-050) supervised by SAR. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Conflict of interest statement

JW is a statistical editor of the Cochrane Gynaecology and Fertility Group. SR is a statistical editor for Hum Reprod. JW also declares that publishing peer-reviewed articles benefits his career. ALM has received a consultation fee from Merck Serono, Ferring and MSD. OR and CF declare no conflicts of interest.

9.6 References for Chapter 9.

Akaike, H. Information theory and an extension of the maximum likelihood principle. Proc. 2nd Int. Symp. Information Theory, Supp. to Problems of Control and Information Theory 1972: 267-281.

Allegra A, Marino A, Volpes A, Coffaro F, Scaglione P, Gullo S, La Marca A. A randomized controlled trial investigating the use of a predictive nomogram for the selection of the FSH starting dose in IVF/ICSI cycles. *Reprod Biomed Online* 2017.

Altman DG, Royston P. The cost of dichotomising continuous variables. *Bmj* 2006;332: 1080.

Arce JC, Andersen AN, Fernandez-Sanchez M, Visnova H, Bosch E, Garcia-Velasco JA, Barri P, De Sutter P, Klein BM, Fauser BCJM. Ovarian response to recombinant human follicle-stimulating hormone: a randomized, antimullerian hormone-stratified, dose-response trial in women undergoing in vitro fertilization/intracytoplasmic sperm injection. *Fertil Steril* 2014;102: 1633-U1456.

Broer SL, Dolleman M, Opmeer BC, Fauser BC, Mol BW, Broekmans FJM. AMH and AFC as predictors of excessive response in controlled ovarian hyperstimulation: a meta-analysis. *Hum Reprod Update* 2011;17: 46-54.

Broer SL, Dolleman M, van Disseldorp J, Broeze KA, Opmeer BC, Bossuyt PMM, Eijkemans MJC, Mol BW, Broekmans FJM, Grp I-ES. Prediction of an excessive response in in vitro fertilization from patient characteristics and ovarian reserve tests and comparison in subgroups: an individual patient data meta-analysis. *Fertil Steril* 2013;100: 420-+.

Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj* 2011;343: d5928.

Klinkert E, Velde ET, Broekmans F. 'Defining poor ovarian response during IVF cycles, in women aged < 40 years, and its relationship with treatment outcome'. *Hum Reprod* 2005;20: 573-573.

Kurinczuk JJH, C. Fertility Treatment in 2006 - a statistical analysis. *Human Fertilisation and Embryology Authority* 2010; London.

La Marca A, Papaleo E, Grisendi V, Argento C, Giulini S, Volpe A. Development of a nomogram based on markers of ovarian reserve for the individualisation of the follicle-stimulating hormone starting dose in in vitro fertilisation cycles. *BJOG* 2012;119: 1171-1179.

La Marca A, Sunkara SK. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Hum Reprod Update* 2014;20: 124-140.

Lekamge DN, Lane M, Gilchrist RB, Tremellen KP. Increased gonadotrophin stimulation does not improve IVF outcomes in patients with predicted poor ovarian reserve. *J Assist Reprod Genet* 2008;25: 515-521.

McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* 2013;32: 3388-3414.

Nelson SM, Yates RW, Fleming R. Serum anti-Mullerian hormone and FSH: prediction of live birth and extremes of response in stimulated cycles - implications for individualization of therapy. *Hum Reprod* 2007;22: 2414-2421.

Nelson SM, Yates RW, Lyall H, Jamieson M, Traynor I, Gaudoin M, Mitchell P, Ambrose P, Fleming R. Anti-Mullerian hormone-based approach to controlled ovarian stimulation for assisted conception. *Hum Reprod* 2009;24: 867-875.

Nyboe Andersen A, Nelson SM, Fauser BC, Garcia-Velasco JA, Klein BM, Arce JC, group E-s. Individualized versus conventional ovarian stimulation for in vitro fertilization: a multicenter, randomized, controlled, assessor-blinded, phase 3 noninferiority trial. *Fertil Steril* 2017;107: 387-396 e384.

Popovic-Todorovic B, Loft A, Bredkjaeer HE, Bangsboll S, Nielsen IK, Andersen AN. A prospective randomized clinical trial comparing an individual dose of recombinant FSH based on predictive factors versus a 'standard' dose of 150 IU/day in 'standard' patients undergoing IVF/ICSI treatment. *Hum Reprod* 2003;18: 2275-2282.

R Core Team. R: A language and environment for statistical computing. 2014. R Foundation for Statistical Computing, Vienna, Austria.

Rustamov O, Smith A, Roberts SA, Yates AP, Fitzgerald C, Krishnan M, Nardo LG, Pemberton PW. Anti-Mullerian hormone: poor assay reproducibility in a large cohort of subjects suggests sample instability. *Hum Reprod* 2012;27: 3085-3091.

Schultz BB. Levene Test for Relative Variation. *Syst Zool* 1985;34: 449-456.
Senn S. Mastering variation: variance components and personalised medicine. *Stat Med* 2016;35: 966-977.

Snijders TAB, Bosker RJ. *Multilevel analysis : an introduction to basic and advanced multilevel modeling*. 2nd edn, 2012. SAGE, Los Angeles ; London.

Stan Development Team. RStan: the R interface to Stan, Version 2.5.0. 2014.

Sterne JAC, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *Bmj-Brit Med J* 2016;355.

Sterrenburg MD, Veltman-Verhulst SM, Eijkemans MJC, Hughes EG, Macklon NS, Broekmans FJ, Fauser BCJM. Clinical outcomes in relation to the daily dose of recombinant follicle-stimulating hormone for ovarian stimulation in in vitro fertilization in presumed normal responders younger than 39 years: a meta-analysis. *Hum Reprod Update* 2011;17: 184-196.

Steward RG, Lan L, Shah AA, Yeh JS, Price TM, Goldfarb JM, Muasher SJ. Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: an analysis of 256,381 in vitro fertilization cycles. *Fertil Steril* 2014;101: 967-973.

Su YS, Gelman A, Hill J, Yajima M. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *J Stat Softw* 2011;45: 1-31.

Sunkara SK, La Marca A, Seed PT, Khalaf Y. Increased risk of preterm birth and low birthweight with very high number of oocytes following IVF: an analysis of 65 868 singleton live birth outcomes. *Hum Reprod* 2015;30: 1473-1480.

Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. *Hum Reprod* 2011;26: 1768-1774.

van Tilborg TC, Broekmans FJM, Dolleman M, Eijkemans MJC, Mol B, Laven JSE, Torrance HL. Individualized follicle-stimulating hormone dosing and in vitro fertilization outcome in agonist downregulated cycles: a systematic review. *Acta Obstet Gynecol Scand* 2016;95: 1333-1344.

Walker AM. Confounding by indication. *Epidemiology* 1996;7: 335-336.

9.7 **Supplementary Material for Chapter 9.**

[Supplementary material for this chapter begins on the next page]

Protocol 1 (01 Sep 2008-31 Dec 2010)	**Protocol 2 (V1)** (01 Jan 2011-30 Apr 2011)	**Protocol 2 (v2)** (01 May 2011-31 Jul 2011)	**Protocol 2 (v3)** (01 Aug 2011-30 Nov 2011)	**Protocol 2 (v4)** (01 Dec 2011-08 Aug 2012)
Initial dose (Day 1-3)	**Initial dose (Day 1-3)**	**Initial dose (Day 1-3)**	**Initial dose (Day 1-3)**	**Initial dose (Day 1-3)**
1) <2.2 AMH (DSL) *Exclude*	**1) <3 AMH (Gen II)** *Co-Flare: 450 hMG*	**1) <3 AMH (Gen II)** *Co-Flare: 450 hMG*	**1) 2-3 AMH (Gen II)** *Antagonist: 450 hMG*	**1) 2-3 AMH (Gen II)** *Antagonist: 300 rFSH*
2) 2.2-15.6 AMH (DSL) *Antagonist: 300 hMG*	**2) 3-10 AMH (Gen II)** *Antagonist: 375 hMG*	**2) 3-10 AMH (Gen II)** *Antagonist: 300 hMG*	**2) 3-10 AMH (Gen II)** *Long Agonist: 300 hMG*	**2) 3-10 AMH (Gen II)** *Long Agonist: 225 rFSH*
3) 15.7-28.5 AMH (DSL) *Long Agonist***:** *200 rFSH/225 hMG*	**3) 11-21 AMH (Gen II)** *Long Agonist: 300 hMG*	**3) 11-21 AMH (Gen II)** *Long Agonist: 225 hMG*	**3) 11-21 AMH (Gen II)** *Long Agonist: 225 hMG*	**3) 11-21 AMH (Gen II)** *Long Agonist: 187.5 rFSH*
4) >28.6 AMH (DSL) *Antagonist: 150 hMG*	**4) 22-30 AMH (Gen II)** *Long Agonist: 225 hMG*	**4) 22-39 AMH (Gen II)** **without PCOS** *Long Agonist: 150 hMG*	**4) 22-39 AMH (Gen II)** **without PCOS** *Long Agonist: 150 hMG*	**4) 22-39 AMH (Gen II)** **without PCOS** *Long Agonist: 150 hMG*
	5) 31-39 AMH (Gen II) *Long Agonist: 150 hMG*	**5) 22-39 AMH (Gen II)** **with PCOS** *Long Agonist: 150 rFSH*	**5) 22-39 AMH (Gen II)** **with PCOS** *Antagonist: 150 rFSH*	**5) 22-39 AMH (Gen II)** **with PCOS** *Antagonist: 150 hMG*
	6) 40-67 AMH (Gen II) without PCO Long Agonist: 150 hMG	**6) 40-67 AMH (Gen II)** **without PCOS** *Long Agonist: 150 hMG*	**6) 40-67 AMH (Gen II)** without PCOS *Antagonist: 150 hMG*	**6) 40-67 AMH (Gen II)** **without PCOS** *Antagonist: 150 hMG*
	7) 40-67 AMH (Gen II) with PCO *Long Agonist: 125 rFSH*	**7) 40-67 AMH (Gen II)** **with PCOS** *Long Agonist: 112.5 rFSH*	**7) 40-67 AMH (Gen II)** **with PCOS** *Antagonist: 112.5 rFSH*	**7) 40-67 AMH (Gen II)** **with PCOS** *Antagonist: 112.5 hMG*
	8) >67 AMH (Gen II) *Long Agonist: 112.5 rFSH*	**8) >67 AMH (Gen II)** *Long Agonist: 112.5 rFSH*	**8) >67 AMH (Gen II)** *Antagonist: 112.5 rFSH*	**8) >67 AMH (Gen II)** *Antagonist: 112.5 hMG*
Dose adjustment *No or minimum change on daily dose of gonadotrophin*	**Dose adjustment** *Step up or down using Oestradiol levels (Day 3&6) and Ultrasound follicle tracking (Day 8)*	**Dose adjustment** *Step up or down using Oestradiol levels (Day 3&6) and Ultrasound follicle tracking (Day 8)*	**Dose adjustment** *Step up or down using Oestradiol levels (Day 3&6) and Ultrasound follicle tracking (Day 8)*	**Dose adjustment** *Step up or down using Oestradiol levels (Day 3&6) and Ultrasound follicle tracking (Day 8)*

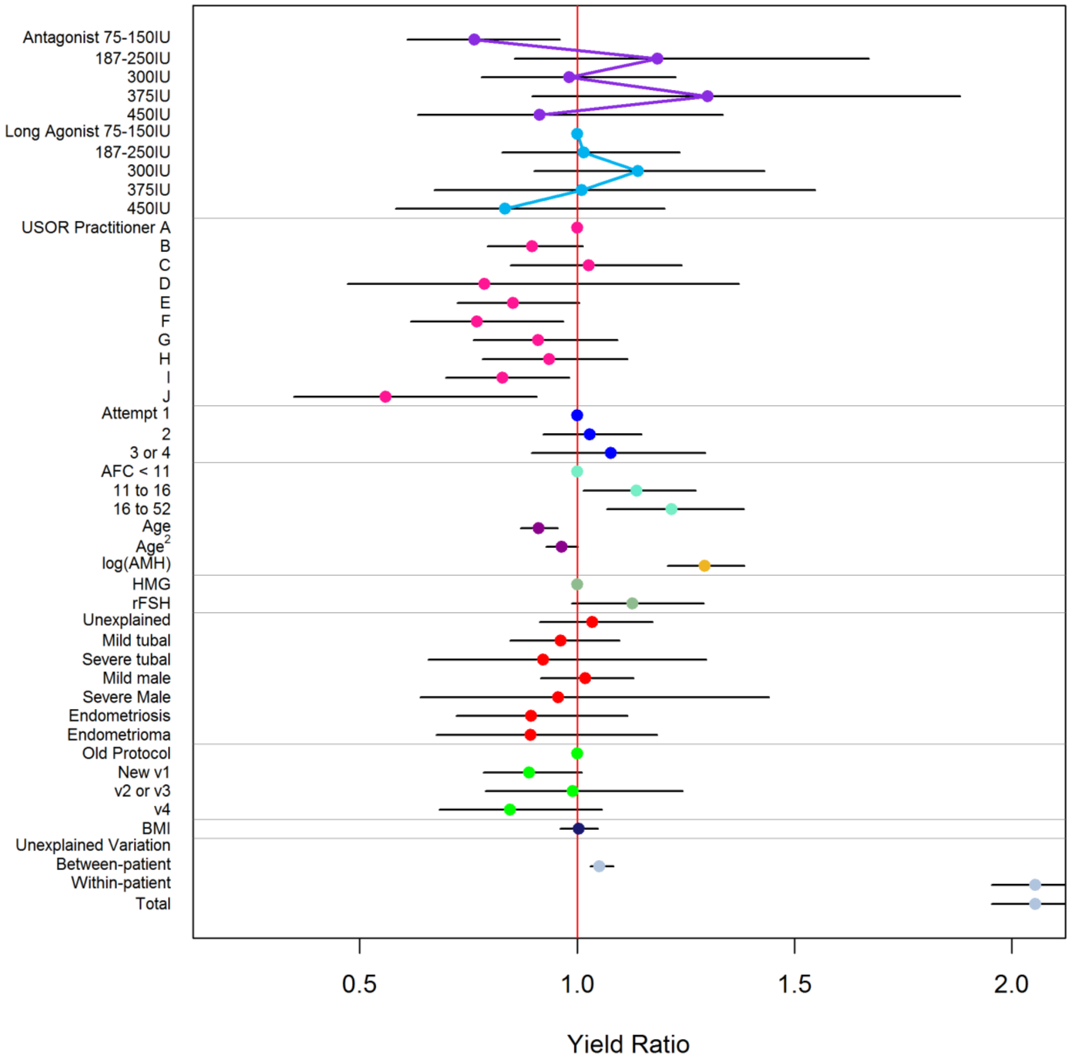
S Table 18: AMH-tailored stratification protocols for regime, starting dose of hMG/rFSH and adjusting daily dose of gonadotrophins (St Mary's Hospital).

	Protocol 1 (01 Sep 2008-31 Dec 2010)	**Protocol 2 (v1)** (01 Jan 2011-30 Apr 2011) & **Protocol 2 (v2)** (01 May 2011-31 Jul 2011)	**Protocol 2 (v3)** (01 Aug 2011-30 Nov 2011)	**Protocol 2 (v4)** (01 Dec 2011-08 Aug 2012)
Coasting for excessive response on day 8	Oestradiol >20,000 pg/ml	30-40 follicles larger than 10mm or Oestradiol >18,000 pg/ml	30-40 follicles larger than 12mm	No coasting
Coasting for excessive response once follicle maturation meets criteria	Oestradiol >20,000 pg/ml	30-40 follicles larger than 10mm	25-40 follicles larger than 10mm	25-30 follicles larger than 15mm
Cancellation for excessive response	Day 8 or thereafter Oestradiol l>20,000 pg/ml and symptoms of OHSS after >3 days of coasting	Day 8 or thereafter More than 40 follicles larger than 10mm	Day 10 or thereafter More than 40 follicles larger than 15mm	Day 8 or thereafter Cancel only if symptoms of OHSS

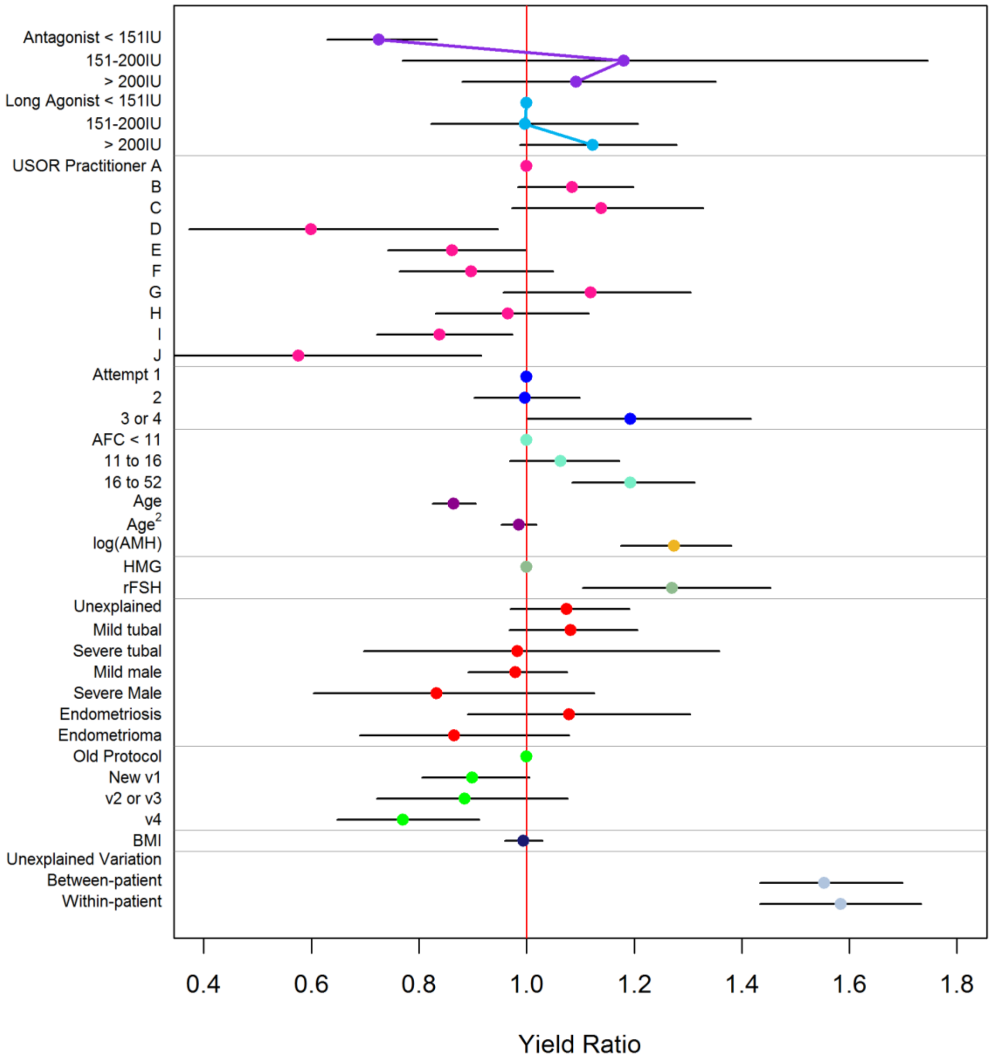
S Table 19: AMH-tailored stratification protocols for management of suspected excessive response (St Mary’s Hospital).

Parameter	**Number of oocytes**	**Number of MII oocytes**
Intercept	8.91 (7.79 to 10.22)	7.14 (5.27 to 9.64)
Treatment characteristics		
Long Agonist 75-150 IU	1.00	1.00
Long Agonist 187-250 IU	1.12 (1.01 to 1.25)	1.02 (0.83 to 1.24)
Long Agonist 300 IU	1.17 (1.03 to 1.33)	1.14 (0.90 to 1.43)
Long Agonist 375 IU	1.18 (0.92 to 1.51)	1.01 (0.67 to 1.55)
Long Agonist 450 IU	1.07 (0.87 to 1.33)	0.83 (0.58 to 1.20)
Antagonist 75-150 IU	0.76 (0.67 to 0.86)	0.76 (0.61 to 0.96)
Antagonist 187 – 250 IU	1.08 (0.90 to 1.30)	1.19 (0.86 to 1.67)
Antagonist 300 IU	1.04 (0.91 to 1.18)	0.98 (0.78 to 1.23)
Antagonist 375 IU	1.11 (0.90 to 1.37)	1.30 (0.90 to 1.88)
Antagonist 450 IU	0.94 (0.76 to 1.17)	0.91 (0.63 to 1.33)
OPU operator: A	1.00	1.00
B	0.98 (0.91 to 1.04)	0.90 (0.79 to 1.01)
C	1.04 (0.94 to 1.16)	1.03 (0.85 to 1.24)
D	0.68 (0.51 to 0.89)	0.79 (0.47 to 1.37)
E	0.78 (0.71 to 0.86)	0.85 (0.73 to 1.00)
F	0.86 (0.78 to 0.97)	0.77 (0.62 to 0.97)
G	0.95 (0.87 to 1.05)	0.91 (0.76 to 1.09)
H	0.93 (0.84 to 1.02)	0.94 (0.78 to 1.12)
I	0.77 (0.70 to 0.84)	0.83 (0.70 to 0.98)
J	0.70 (0.56 to 0.88)	0.56 (0.35 to 0.91)
Protocol: Old	1.00	1.00
New protocol (V1)	0.87 (0.81 to 0.93)	0.89 (0.79 to 1.01)
New protocol (V2 & V3)	0.90 (0.79 to 1.02)	0.99 (0.79 to 1.24)
New protocol (V4)	0.84 (0.74 to 0.94)	0.85 (0.68 to 1.06)
Patient characteristics		
Attempt No: 1st	1.00	1.00
2nd	1.05 (0.99 to 1.11)	1.03 (0.92 to 1.15)
3rd or 4th	1.19 (1.07 to 1.32)	1.08 (0.90 to 1.29)
Antral follicle count: < 10	1.00	1.00
11 to 16	1.16 (1.11 to 1.23)	1.14 (1.01 to 1.27)
16 to 52	1.29 (1.20 to 1.38)	1.22 (1.07 to 1.38)
Age (SDs)	0.87 (0.85 to 0.89)	0.91 (0.87 to 0.96)
Age^2 (SDs)	0.96 (0.94 to 0.99)	0.96 (0.93 to 1.00)
Log(AMH) (SDs)	1.35 (1.30 to 1.40)	1.29 (1.21 to 1.38)
Gonadotropin: HMG	1.00	1.00
rFSH	1.15 (1.07 to 1.24)	1.13 (0.99 to 1.29)
Unexplained fertility	1.07 (1.00 to 1.14)	1.03 (0.91 to 1.17)
Mild tubal	1.01 (0.94 to 1.08)	0.96 (0.85 to 1.10)
Severe tubal	0.92 (0.77 to 1.09)	0.92 (0.66 to 1.30)
Mild male factor	0.99 (0.93 to 1.05)	1.02 (0.92 to 1.13)
Severe male factor	1.11 (0.88 to 1.40)	0.96 (0.64 to 1.44)
Endometriosis	0.94 (0.85 to 1.06)	0.89 (0.72 to 1.12)
Endometrioma	0.87 (0.75 to 1.02)	0.89 (0.68 to 1.18)
BMI (SDs)	1.01 (0.99 to 1.04)	1.00 (0.96 to 1.05)

S Table 20: Yield ratios and 95% CIs from fitted Poisson regression models of number of oocytes and of number of mature oocytes, with the covariates shown in the table. Estimates for treatment characteristics relate to total effects after holding patient characteristics fixed. Estimates for patient characteristics relate to direct effects on COS response (ie: after subtracting the 'effect' of characteristics on treatment selection).



S Figure 1: Yield ratios and 95% CIs from the multivariable Poisson regression model of number of metaphase II oocytes per cycle. Continuous predictors have been standardized, so that coefficients display the expected multiplicative increase in the yield ratio for a standard deviation change in the variable. Increasing dose effect under a GnRH Antagonist regime is shown by the purple connecting line. Increasing dose effect under a GnRH Long Agonist regime is shown by the blue connecting line



S Figure 2: Sensitivity analysis. Yield ratios and 95% CIs from a multivariable Poisson regression model of number of oocytes per cycle, restricted to low gonadotropin doses. Continuous predictors have been standardized, so that coefficients display the expected multiplicative increase in the yield ratio for a standard deviation change in the variable. Increasing dose effect under a GnRH Antagonist regime is shown by the purple connecting line. Increasing dose effect under a GnRH Long Agonist regime is shown by the blue connecting line.

Chapter 10. Analysis of multistage in vitro fertilization data with mixed multilevel outcomes using joint modelling approaches

Journal article 5

Authors Jack Wilkinson, Andy Vail, Stephen A Roberts

Status Under peer review at Journal of the Royal Statistical Society Series C

Contribution statement JW devised the manuscript, prepared the dataset, conducted and interpreted statistical analysis and co-authored the manuscript. SR and AV devised the manuscript, contributed to the interpretation of statistical analysis and co-authored the manuscript.

Preamble In this paper we present three different approaches to the analysis of multistage IVF data (the correlated latent variable approach, the endogenous response approach, and outcome regression, where we fit a series of conditional regression models for each stage of the cycle). We describe the features of each approach before applying all three methods to routine clinical data in a relatively simple example. The target audience for this article is applied statisticians, so we provide both a mathematical description of the model, as well as Stan code.

Outputs and Impact of the research A version of the manuscript is currently under review. The work has been presented as oral presentations at the International Society for Clinical Biostatistics Conference 2017 and at the Statistical Analysis of Multi-Outcome Data Meeting 2017, where it was awarded the prize for best presentation. A preprint posted at bioRxiv (BIORXIV/2017/173534) has been downloaded 62 times at the time of writing.

10.1 Abstract

In vitro fertilization comprises a sequence of interventions concerned with the creation and culture of embryos which are then transferred to the patient's uterus. While the clinically important endpoint is birth, the responses to each stage of treatment contain additional information about the reasons for success or failure. Joint analysis of the sequential responses is complicated by mixed outcome types defined at two levels (patient and embryo). We develop three methods for multistage analysis based on joining submodels for the different responses using latent variables and entering outcome variables as covariates for downstream responses. An application to routinely collected data is presented, and the strengths and limitations of each method are discussed.

Keywords: in vitro fertilisation; joint modelling; mixed data; multilevel modelling; multistage treatment data; multivariate responses

10.2 **Background and motivation**

In vitro fertilization (IVF) is a complex multistage procedure for the treatment of subfertility. Typically, a ‘cycle’ of IVF begins with the administration of drugs to stimulate the patient’s ovaries and promote the release of oocytes (eggs). The oocytes are collected from the patient and are then fertilised either by mixing or injecting them with sperm. The resulting embryos are cultured for several days. Finally, one or more of the best embryos are selected for transfer to the woman’s uterus, where it is hoped that they will implant and develop into a healthy baby. Treatment may fail at any stage of the cycle (if no oocytes are recovered from the ovaries, no good quality embryos are produced, or those transferred do not implant), in which case the subsequent stages are not undertaken.

The sequential nature of IVF means that the patient’s response can be measured at each stage of the treatment (Heijnen et al., 2004): the stimulation of the ovaries can be evaluated by the number of oocytes collected; the fertilization and culture stages can be evaluated by the number and quality of embryos produced; and the success of the transfer procedure can be evaluated according to whether or not a child is born as a result. Figure 37 displays a schematic of the IVF cycle. A recent review of outcome measures used in IVF RCTs showed that there is considerable interest in these ‘intermediate’ or ‘procedural’ outcomes of IVF; 361 distinct numerators were identified, and the median (IQR) number of distinct outcomes reported per trial was 11 (7 to 16) (Wilkinson et al., 2016).

The interest in procedural outcomes in IVF research is not surprising. While the most relevant measure of success for patients is the birth of a child (Min et al. 2004, Heijnen et al. 2004, Legro et al. 2014), establishing the effects of treatments and patient characteristics on procedural outcomes might increase our mechanistic understanding of how IVF works and how it might be improved. The question of how outcomes at each stage of the process relate to one another also appears to be relevant to designing and evaluating IVF interventions. In response, two approaches for the analysis of multistage IVF data have recently been proposed (Maity et al. 2014, Penman et al. 2007). The first is a discrete time-to-event approach that treats the stages of the IVF cycle as a series of 'failure opportunities' (Maity et al. 2014). Each woman's response data then comprise a vector of binary indicator variables denoting whether they failed at this stage, or proceeded to the next. The second treats the stage of the cycle reached by the patient as an ordinal response, and models this using continuation ratio regression (Penman et al.

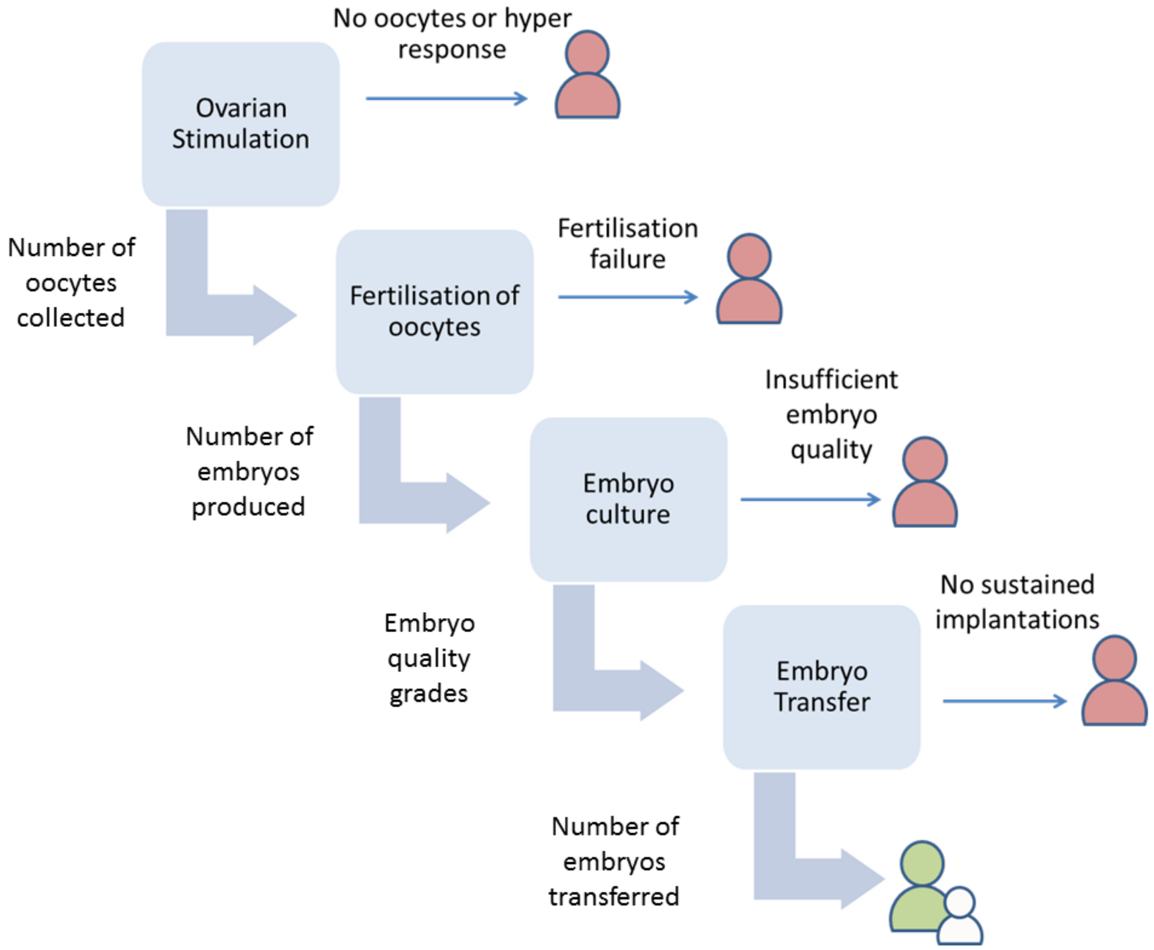

Ovarian Stimulation
No oocytes or hyper response
Number of oocytes collected
Fertilisation of oocytes
Fertilisation failure
Number of embryos produced
Embryo culture
Insufficient embryo quality
Embryo quality grades
Embryo Transfer
No sustained implantations
Number of embryos transferred

Figure 37: Schematic of the IVF cycle.

2007). Both of these approaches allow us to answer research questions relating to the effects of baseline treatment and patient characteristics on IVF response, while preserving the sequential nature of the data. Both share similar limitations, however. In particular, both treat the responses at each stage as dichotomous 'success or failure' events. This wastes a great deal of information, since it is more informative to measure the number of oocytes obtained from the ovaries than merely whether a sufficient quantity were available to enable the cycle to continue; and it is more informative to measure the quality of any embryos obtained than merely whether there were any available for transfer. These methods are also incapable of accommodating outcomes defined at different levels of a multilevel structure; some outcomes (eg: number of oocytes) may be defined for each patient, while others (eg: embryo quality) are defined for the patient's individual embryos. In addition, while these methods allow for differential effects of covariates at each stage through the inclusion of interaction terms, they do not allow for different covariates to be included as predictors for the different stage-specific responses.

While methods for the analysis of sequential IVF data exist therefore, it remains to identify techniques capable of incorporating the variety of outcome types encountered in this setting, and moreover responses which are defined at different levels of a two-level data structure (embryos and patients). This includes counts of oocytes, ordinal embryo quality scales, binary birth indicator variables, and so on. Methods for the analysis of multivariate responses of mixed outcome types are hardly new (eg: Goldstein 2003) but have received considerable attention in recent years (see de De Leon and Chough, 2013 for a comprehensive collection of the state of the art). While much of this work has focussed on the joint analysis of time-to-event and longitudinal response data (see reviews by Gould et al., 2015 and Tsiatis and Davidian, 2004), approaches capable of accommodating different combinations of outcome types have been described (McCulloch, 2008, Dunson, 2000, Gueorguieva, 2001, Gueorguieva and Agresti, 2001, Dunson and Herring, 2005, Dunson et al., 2003, Goldstein et al., 2009). Typically, these involve the inclusion of shared (McCulloch 2008, Dunson 2000, Gueorguieva 2001, Dunson, 2000, Dunson and Herring 2005) or otherwise correlated (Gueorguieva and Agresti 2001, Goldstein et al., 2009) latent variables in 'submodels' for the different response variables. These latent variables accommodate dependency between the response variables in the model. Moreover, by estimating the parameters governing the distribution of these latent variables, we can examine both the direction and degree of association between a patient's responses. A further attractive feature of latent variable approaches is that they can be used to jointly model responses measured at different levels of a multilevel data structure (Goldstein et al. 2009, Dunson et al., 2003). These methods do not appear to have been discussed in the context of multistage treatments however.

Given the strict temporal ordering of the stages, an alternative strategy for the analysis of IVF data would be to explicitly model the relationships between the patient's stage-specific responses using a series of conditional regression equations (Blalock, 1961). Under this sort of approach, each response variable would be included as a covariate in the regression equations for each of the subsequent, or 'downstream', responses. An advantage of these approaches is that they allow direct and indirect effects of the procedural responses on downstream outcomes to be distinguished (Pearl, 2001). A third strategy we might consider would be to combine the two approaches hitherto described, and simultaneously link submodels for each response using latent variables while including the response variables as covariates in the downstream response models. This would then resemble the endogeneous treatment models employed in the econometrics literature (Terza, 1998), or multiprocess models that have been employed in education research (Steele et al., 2009).

In this paper, we develop methodology for the analysis of multistage IVF data, with mixed response types (count, ordinal, and binary) defined at different levels of a two-level data structure (patients and embryos). We describe three approaches in which distinct submodels are used for the various response variables. In the first, we include correlated latent variables and estimate the relationships between the responses. In the second, we adopt an outcome regression approach where response variables enter into regression equations for downstream response variables as covariates. This approach can be considered as a set of separate regression models. In the third, we consider an endogenous response model where we combine both of these approaches, by including upstream response variables as covariates in downstream submodels, and also allowing the submodel-specific latent variables to be correlated. The remainder of the manuscript is structured as follows. In section 10.3, we describe the models. In section 10.4 we illustrate the use of the methods with an application to a routine clinical database. This is followed by a discussion in section 10.5. We conclude with some brief recommendations in section 10.6.

10.3 **Models**

Here we describe latent variable, outcome regression and endogenous response modelling approaches to the analysis of multistage IVF data. The approaches have several key features in common. First, they all include distinct submodels for each of the response variables considered in the cycle. We include six response variables for patient *j* = 1,…,*n* and their embryos *i* = 1,…,n_j, and hence six submodels, in the current presentation: the number of oocytes (eggs) obtained from ovarian stimulation (a count, y_j^O); the fertilisation rate when the oocytes are mixed with sperm (y_j^M); two measures of embryo quality (cell evenness and degree of fragmentation $y_{ij}^E$and y_{ij}^F, both measured using ordinal grading scales); an indicator denoting whether one or two embryos were transferred to the patient (denoted by a binary variable y_j^D) and another (y_j^L) indicating whether or not the transfer of embryos resulted in the live birth of one or more babies (a live birth event, or LBE) (Figure 37). These are listed in temporal order, with the exception of the two embryo quality scales, which are coincident. We include the decision to transfer two embryos (known as double embryo transfer, or DET) in the model because it is an important predictor of transfer success which is partially determined by the outcomes of the earlier stages. A second feature common to the approaches is that once a patient has dropped out of the cycle, they do not appear in the submodels corresponding to the downstream responses. In the following, we ignore the possibility that each patient may undergo multiple cycles of IVF, noting that the models could be extended to three levels (embryos nested within cycles nested within women) by adding additional random scalar terms (Goldstein, 2011).

10.3.1. **Correlated latent variable approach**

This approach requires the use of latent variable representations for the various submodels constituting the larger model. Each patient *j* has associated vectors of responses $\boldsymbol{Y}_j$= $(y_j^O, y_j^M, y_{ij}^E, y_{ij}^F, y_j^D, y_j^L)$ and of underlying latent variables $\boldsymbol{Z}_\mathrm{j}$ = $(z_j^O, z_j^M, z_j^E, z_j^F, z_j^D, z_j^L)$. Both of these vectors may be partially observed due to drop-out or outright failure before completion of the treatment. We then posit a multivariate Normal distribution for the latent variables, and estimate the elements of the correlation and variance-covariance matrices. We prefer to use distinct latent variables in each submodel to an approach based on a common latent variable which is scaled by factor loadings in each submodel (eg: Dunson, 2000, McCulloch, 2008), as the linearity assumption required for the latter is too restrictive for present purposes (Gueorguieva, 2001). For the latent variable approach, we do not include response variables as covariates in any of the submodels. The submodels for each stage are presented below, followed by the multivariate distribution of latent variables.

Stimulation phase submodel

For patient *j,* we assume the number of oocytes (eggs) obtained y_j^O follows a Poisson distribution and model the log of the rate parameter λ_j^o in the usual way:

$$\log(\lambda_j^o) = \boldsymbol{X}_j^o \boldsymbol{\beta}^o + z_j^o \qquad (1)$$

where $\boldsymbol{X}_j^o$ is a row-vector of cycle-level covariates for patient *j*, $\boldsymbol{\beta}^o$ is a corresponding vector of regression parameters and z_j^o is a patient-specific latent variable that captures overdispersion in the oocyte yield. This submodel is fitted to all patients who start the cycle.

Fertilisation submodel

We model the number of embryos obtained when oocytes are mixed with sperm y_j^M in terms of its rate parameter λ_j^M, again using a Poisson submodel:

$$\log(\lambda_j^M) = \log(y_j^O) + \boldsymbol{X}_j^M \boldsymbol{\beta}^M + z_j^M \qquad (2)$$

where $\boldsymbol{X}_j^M$, $\boldsymbol{\beta}^M$ and z_j^M are analogous to the corresponding terms in the stimulation model. We now include an offset term corresponding to the logarithm of the number of oocytes obtained in the linear predictor. This submodel is fitted to all patients who have oocytes mixed with sperm. In some cycles, the number of oocytes mixed with sperm is less than the number obtained, so there is an implicit assumption in the model that any oocytes which were not mixed could not have been successfully fertilized. The assumption is reasonable, since the decision not to mix an oocyte with sperm is almost always based on the fact that the oocyte has been identified as being degenerate.

Embryo quality submodels

We include two measures of embryo quality; cell evenness (y^E) and degree of fragmentation (y^F). These are ordinal 1 to 4 grading scales measured at the level of individual embryos. We model these using cumulative logit submodels. For embryo *i* (where *i* = 1,2,…,n_j) nested in patient *j* we have, for *k* = 1,2,3:

$$\begin{aligned} \text{logit}(\gamma_{kij}^E) &= \alpha_k^E - \boldsymbol{X}_{ij}^E \boldsymbol{\beta}_k^E - z_j^E \\ \text{logit}(\gamma_{kij}^F) &= \alpha_k^F - \boldsymbol{X}_{ij}^F \boldsymbol{\beta}_k^F - z_j^F \end{aligned} \tag{3}$$

where $\boldsymbol{X}_{ij}^E$ and $\boldsymbol{X}_{ij}^F$ are row-vectors of covariates, $\boldsymbol{\beta}_k^E$ and $\boldsymbol{\beta}_k^F$ are vectors of regression coefficients which may vary across the levels of *k* (relaxing the proportional odds assumption), and z_j^E and z_j^F are patient-level random effects (latent variables) which are identified due to the clustering of embryos within patients. γ_{kij}^E and γ_{kij}^F are cumulative probabilities of embryo *i* in patient *j* having a grade of *k* or lower for evenness and fragmentation degree respectively and α_k^E and α_k^F are threshold parameters, corresponding to the log-odds of the embryo having grade *k* or lower. These submodels are fitted to all embryos.

Double embryo transfer submodel

In order to jointly model the binary response DET, denoting the number of embryos transferred, with the other response variables, we use a latent variable representation of a probit regression model (Albert & Chib 1993). Let $y_j^D = 1$ or 0 if patient *j* does or does not have DET, respectively. We define y_j^{D*} as a latent continuous variable underlying the binary y_j^D, such that:

$$y_j^D = \begin{cases} 1 \; if y_j^{D*} \;\; \geq 0 \\ 0 \; if \; y_j^{D*} \; < 0 \end{cases} \tag{4}$$

A linear regression submodel for the latent y_j^{D*} is then used to estimate covariate effects:

$$y_j^{D*} \; = \; \boldsymbol{X}_j^D \boldsymbol{\beta}^D + \; z_j^D \tag{5}$$

$$z_j^D \sim N(0, 1)$$

where $\boldsymbol{X}_j^D$ is a row-vector of patient-level covariates and $\boldsymbol{\beta}^D$ is a vector of regression coefficients. Fixing the variance of z_j^D to be 1 is mathematically equivalent to specifying a probit model for the probability that a patient will have DET. We use this error term to link the DET submodel to the others.

Live birth event submodel

As for DET, we use a latent probit representation for $y_j^L = 1$ or 0 corresponding to whether or not LBE obtains, with an underlying latent variable y_j^{L*}, row vector of patient-level covariates $\boldsymbol{X}_j^L$ and vector of regression coefficients $\boldsymbol{\beta}^L$. The error term z_j^L again has a variance of 1, and is used to link the LBE submodel to the others. The DET and LBE submodels are fitted to patients who undergo the transfer procedure.

Covariates in the latent variable approach

An essential feature of the latent variable method is that none of the covariate vectors $\boldsymbol{X}_j^O, \boldsymbol{X}_j^M, \boldsymbol{X}_{ij}^E,\ \boldsymbol{X}_{ij}^F, \boldsymbol{X}_j^D,\ \boldsymbol{X}_j^L$ include any of the response variables in $\boldsymbol{Y}_j$.

Latent variable distribution

We specify a multivariate Normal distribution for the latent variables to connect the submodels:

$$\begin{bmatrix} z_j^O \\ z_j^M \\ z_j^E \\ z_j^F \\ z_j^D \\ z_j^L \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \theta_O^2 & \eta_1\theta_O\theta_M & \eta_2\theta_O\theta_E & \eta_3\theta_O\theta_F & \eta_4\theta_O & \eta_5\theta_O \\ \cdot & \theta_M^2 & \eta_6\theta_M\theta_E & \eta_7\theta_M\theta_F & \eta_8\theta_M & \eta_9\theta_M \\ . & . & \theta_E^2 & \eta_{10}\theta_E\theta_F & \eta_{11}\theta_E & \eta_{12}\theta_E \\ . & . & . & \theta_F^2 & \eta_{13}\theta_F & \eta_{14}\theta_F \\ . & . & . & . & 1 & \eta_{15} \\ \eta_5\theta_O & \ldots & \ldots & \ldots & \ldots & 1 \end{bmatrix}\right) \qquad (6)$$

We note that this framework allows us to estimate the relationships between patient and embryo-level responses.

10.3.2. Outcome regression approach

In the outcome regression approach, we fit the submodels presented for the latent variable method as separate regression models, such that we replace the covariance matrix (6) by a diagnonal matrix. By contrast to the latent variable approach, we now include the response variables in the linear predictors for the downstream submodels. In particular: we include the numbers of oocytes and embryos obtained as covariates in the submodels relating to embryo quality, DET and LBE; we include aggregated measures of embryo quality gradings as covariates in the DET and LBE submodels; and we include DET as a covariate in the LBE submodel. We use the shorthand ‘outcome-covariate’ to refer to instances of response variables appearing as covariates in downstream submodels. While the simplicity of fitting separate regression models makes this an attractive strategy, a weakness of this approach is that it rests on the standard regression assumption that covariates are not *endogenous*, which is to say that they are not correlated with the error term in the submodel. This assumption is unlikely to hold if we include outcome-covariates, due to the likelihood that the different response variables in the submodels share unmeasured predictors.

10.3.3. Combining the latent variable and outcome-regression approaches: an endogenous response model.

A third approach we consider is a combination of the two approaches described above. We represent each response variable using a conditional regression equation including upstream response variables in the linear predictor, as for the outcome regression approach (section 10.3.2). However, we also allow the submodels to be joined through multivariate Normal latent variables as for the correlated latent variable method (section 10.3.1). We estimate the variance-covariance matrix of this distribution, together with the regression parameters. This approach allows for the endogeneity of outcome-covariates, since the correlation between response variables is incorporated through the latent variables (Heckman, 1978, Terza, 1998). Consequently, this approach allows for valid estimation of the effects of upstream upon downstream response variables (Skrondal and Rabe-Hesketh, 2004). Identifying endogenous response models can be challenging however (McConnell et al., 2008, Diggle et al., 2007). Standard strategies include fixing parameters in the model (for example, fixing elements of the latent correlation matrix to be zero) and including instrumental variables in some of the submodels (Xie 2000, Steele et al., 2009; Terza, 2009). These variables should be strongly correlated with the response variable of the submodel in which they appear, but should not otherwise be associated with downstream responses.

10.4 Application of the methods to routinely collected IVF data.

10.4.1. St Mary's Data

We utilise the three methods in an application to a routine clinical database from St Mary's Hospital Department of Reproductive Medicine, Manchester, England. Our aim was to establish whether the endogenous response model would allow us to infer more about the internal structure of the IVF cycle compared to the simpler latent variable and outcome regression methods. The dataset includes 2962 initiated IVF treatments undertaken by 2453 women between 2013 and 2015, including quality data on 12,911 embryos. For present purposes, we ignore the fact that some women underwent multiple cycles, noting that the current models could be extended to a three-level setting (Goldstein, 2011). Characteristics of the treatment cycles in the dataset are presented in Table 15. We include age and partner age in all of the submodels. We standardise these by subtracting the mean value and dividing by a standard deviation. The models are flexible enough to allow different covariates to be included in different submodels; we include attempt number in the number of oocytes and DET submodels, pooling 4th and 5th attempts due to small numbers in these categories. In the embryo evenness and fragmentation submodels, we also include an indicator variable denoting whether the egg was fertilized by injecting it with sperm, or by mixing *in vitro*. We suppose that covariate effects are constant across the levels of the ordinal embryo responses (proportional odds), although the methods can accommodate non-proportionality. We fit three models as described in section 10.3 (correlated latent variable model, outcome regression model, and endogenous response model). Figure 38 shows path diagrams for each of these. Note that we do not distinguish between linear and nonlinear relationships in this diagram.

Variable	Summary
No of cycles started	2962
No of cycles where eggs mixed with sperm	2861
No of gradable embryos	12911
Number of embryo transfer procedures	2501
Age (years)	33
	30 to 36
	21 to 43
Partner Age (years)	35
	32 to 39
	19 to 72
Attempt Number	
1	2132 (72%)
2	659 (22%)
3	147 (5%)
4	4 (0%)
5	20 (0%)
Number of eggs obtained per cycle started	9
	5 to 13
	0 to 50
Number of gradable embryos per cycle started	3
	1 to 5
	0 to 19
Number of embryos transferred per transfer procedure	
1	1049 (42%)
2	1452 (58%)
Live birth event per transfer procedure	
No	1692 (68%)
Yes	809 (32%)

Table 15: Characteristics of the clinical dataset analysed in 10.4. Median, interquartile range and range for continuous variables.

10.4.2. Fitting the models

We use the R (R Core Team, 2014) implementation of the Bayesian software Stan (Stan Development Team, 2014) to fit the models. While the benefits (or drawbacks, depending on one's perspective) of Bayesian methods have been well rehearsed, our use of this software is primarily driven by pragmatism; the software is flexible and can accommodate complex multilevel models without the need to author custom sampling algorithms. We place weak Normal (0,1000^2) priors on the regression parameters in the submodels, with the exception of those included in the latent probit submodels (that is, those corresponding to DET, LBE). Given the fact that the latent responses in these submodels have a variance of 1, we place Normal (0, 22) priors on the regression parameters. These can be considered to be weakly informative prior distributions which improve efficiency in fitting the model by restricting the sampler to realistic values for these parameters (Gelman et al., 2014). We place weakly informative Cauchy (0,2.5) priors on the free variance parameters. Finally, we use an LKJ prior distribution for the latent correlation matrix, which is uniform over all possible correlation matrices (Stan Development Team, 2017). We consider this appropriate given that estimation of this matrix is of particular interest in our latent variable models. We run the samplers for between two and three thousand iterations in each case, using three chains. We check convergence using the Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992) and using traceplots. In practice, we note that fitting the endogenous response model can take around 12 hours on an Intel Core i7-4810MQ 2.8 GHz processor with 16 GB of RAM. Stan code is provided in the online supplement available at https://www.biorxiv.org/content/early/2017/08/10/173534.

10.4.3. Results and interpretation

The models produce a large number of parameter estimates relating to the covariates in each submodel and the relationships between stage-specific responses. In the following, we focus on the information provided by each approach regarding the relationships between response variables, and simply note here that estimates relating to other covariates were generally similar between the models.

Latent variable approach

In the latent variable approach, information regarding the relationships between the variables is obtained through the estimated latent correlation structure (Table 2). We note that estimates derived from this model are generally consistent with current understanding. For example, the model suggests a positive relationship between embryo evenness on the one hand, and probability of LBE on the other (transferring higher quality embryos makes success more likely), although the association between fragmentation and LBE is less clear. The number of oocytes obtained from ovarian stimulation and fertilization rate also appear to be associated with LBE (reflecting advantages of having a larger pool of embryos from which to select). Upstream measures of success are negatively associated with DET, possibly due to the fact that the transfer of multiple embryos is usually employed to compensate for poor prognosis. On the other hand, it is not immediately obvious why fertilization rate is (quite strongly) negatively related to the number of oocytes obtained, and to embryo quality variables.

More generally, we might ask whether the latent correlations arising from this approach can reasonably be given a causal interpretation. In relation to this, we note that the estimated correlation coefficients can be adjusted for confounding variables by including these in both of the relevant submodels. However, giving correlation coefficients a causal interpretation may be problematic even if they are appropriately adjusted, since their magnitude is in part determined by the variance of the covariables under consideration, which may vary across populations (Greenland et al., 1991). Moreover, the estimated correlation between any two response variables is not adjusted for other response variables in the model. As a result, it is not possible to distinguish genuine from spurious structural relationships attributable to confounding by other outcome variables. Consequently, the latent variable approach does not appear to yield interpretable estimates of the relationships between response variables. In section 4, we suggest that the latent variable approach may be more useful for the purpose of making multivariate predictions regarding IVF cycle outcomes. We note here that the latent variable approach accommodates both multilevel response data and participants with incomplete response

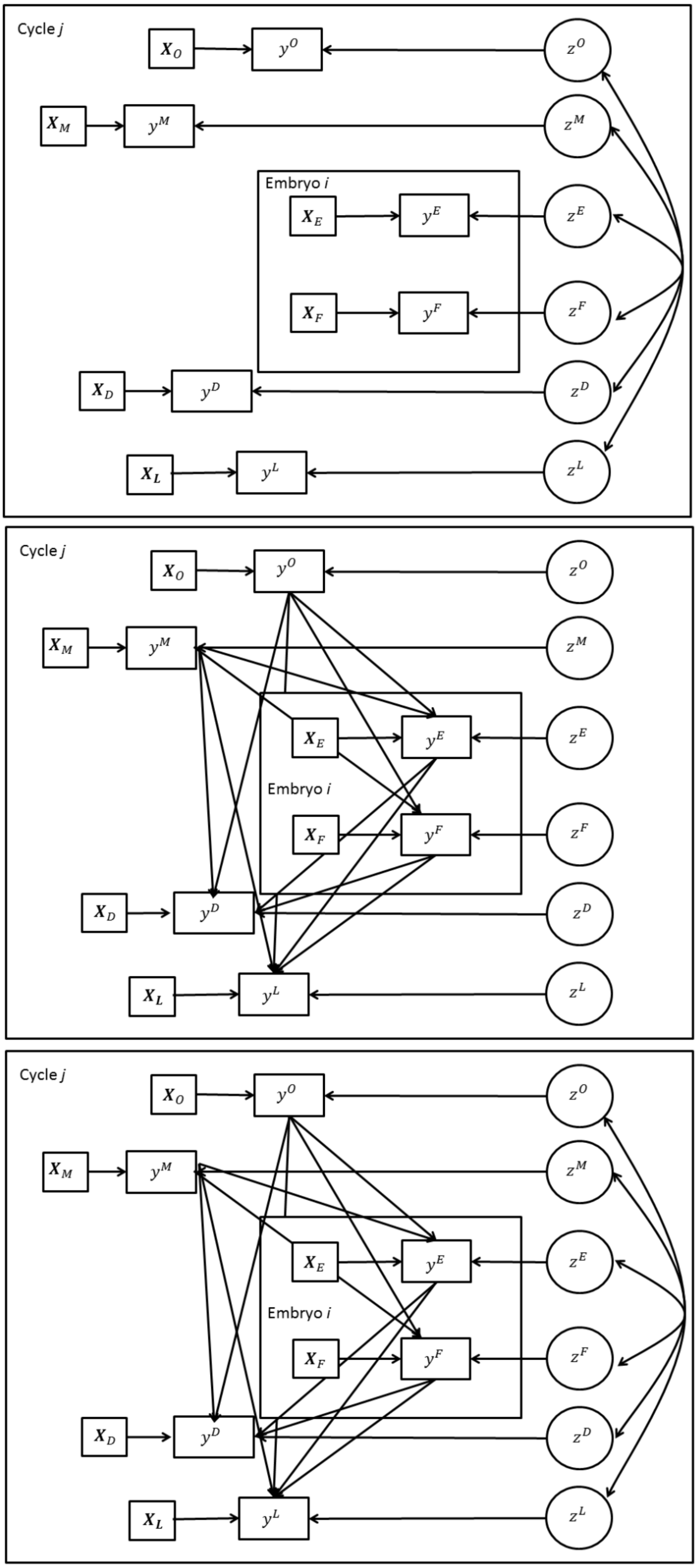


Figure 38: Path diagrams for three models of the IVF cycle: a latent variable model (top), an outcome regression model (middle), and an endogenous response model (bottom).

data. Correlations relating to embryo-level responses can be interpreted as measures of association with the patient's mean values of fragmentation and evenness, while drop out is assumed to be 'missing at random' (MAR, Rubin 1976) given the observed responses and covariate data (McCulloch, 2008).

Outcome regression approach

The outcome regression approach provides information on relationships between response variables directly by way of estimated regression coefficients (Table 17). Unlike the correlation coefficients from the latent variable model (Table 16), these are adjusted for upstream response variables as well as the other covariates in the submodel. The regression coefficients are also easier to interpret compared to the latent correlations and, moreover, may be given a causal interpretation. In the outcome regression approach, the parameter corresponding to fertilization rate in the LBE submodel is an estimate of the effect of increasing fertilization rates on LBE for a fixed number of oocytes, after blocking effects acting via the intermediate outcomes embryo quality or DET (Westreich and Greenland, 2013). The estimate (95% CI) is 0.14 (0.08 to 0.21), indicating a positive effect. For the estimates in the outcome regression model to be valid however, we must assume that there is no unmeasured confounding (Westreich and Greenland, 2013). For example, for our estimates of the effects of embryo evenness and fragmentation in the LBE submodel to be valid, we must assume that there are no unmeasured variables which influence both embryo quality and LBE. This is unrealistic in practice, as there are likely to be deleterious factors which influence both embryo viability and uterine receptivity (Roberts et al., 2010). Even if we believed that this could be adequately accounted for by the inclusion of age as a covariate, residual confounding due to measurement error and model misspecification (for example, including age as a linear term when its relationship with several of the responses may be nonlinear) would ensure that this assumption did not hold (Sterne et al., 2016). In the outcome regression approach, subgroups of participants enter each submodel according to their progress through (and drop out from) the stages of treatment. The model assumes therefore that missing data can be accounted for by the predictor variables in each submodel (and are therefore MAR given these covariates).

Endogenous response modelling approach

The latent correlation matrix from the fitted model is not obviously interpretable (S Table 21). Instead, we use the regression coefficients to investigate the relationships between variables (Table 17). As for the outcome-regression model, the regression coefficient corresponding to an outcome-covariate can be interpreted as an estimate of the effect of the outcome on the response variable in the submodel. This estimate applies for fixed values of any upstream (in relation to our outcome-covariate) variables, after blocking indirect effects through intermediate (downstream) response variables. We note that several of the estimates are inconsistent with those obtained from the outcome regression model. For example, the estimate (95% CI) corresponding to fertilization rate in the LBE submodel changes from 0.14 (0.08 to 0.21) in the outcome regression model to -0.17 (-0.35 to 0.03) in the endogenous response model. This suggests that the positive relationship between fertilization rate and LBE probability observed in the previous models might have been an artefact due to measurement error and unmeasured confounders; in the endogenous response model we conclude that an increased fertilization rate is probably associated with a reduced chance of a successful transfer. This might reflect an increased risk of selecting inferior embryos that are not identified by the grading scales used here. This contrast highlights the possibility of using endogenous response models to explore the extent of unmeasured confounding. In general, the estimates of outcome-covariate effects are less precise in the endogenous response model compared to the outcome regression model. This is analogous to the impact of allowing for, rather than ignoring, clustering of repeated measurements in a mixed model. To check the model, we plotted the observed responses in the data against replicated data drawn from the posterior predictive distribution (Figure 39). For embryo evenness, embryo fragmentation, DET and LBE, we plotted the observed frequency distributions with 95% intervals for the predictive distributions from the model. These checks suggested that the model was compatible with the study data, with the exception of DET, which systematically exceeded the model predictions by a small amount (Figure

39). This is because our prior for the DET intercept was too strong, resulting in underfitting. We would relax this in future applications.

The endogenous response model is again a MAR approach, as missing data are assumed to be ignorable given observed response and covariate data. Since the endogenous response model provides interpretable estimates of effects of procedural responses on downstream events (unlike the latent variable model) while allowing the assumption that outcome-covariates are not endogenous to be relaxed (unlike the outcome regression model) we conclude that this approach is superior for the analysis of multistage IVF treatment data.

	Number of oocytes	Fertilization rate	Embryo evenness	Embryo fragmentation	DET	LBE
Number of oocytes	1	-0.62 (-0.68 to -0.56)	-0.01 (-0.08 to 0.06)	0.03 (-0.03 to 0.09)	-0.09 (-0.16 to -0.02)	0.16 (0.09 to 0.23)
Fertilization rate	-0.62 (-0.68 to -0.56)	1	-0.21 (-0.31 to -0.11)	-0.28 (-0.37 to -0.19)	0.01 (-0.09 to 0.10)	0.11 (0.02 to 0.21)
Embryo evenness	-0.01 (-0.08 to 0.06)	-0.21 (-0.31 to -0.11)	1	0.87 (0.84 to 0.89)	-0.26 (-0.32 to -0.20)	0.06 (0.00 to 0.12)
Embryo fragmentation	0.03 (-0.03 to 0.09)	-0.28 (-0.37 to -0.19)	0.87 (0.84 to 0.89)	1	-0.23 (-0.29 to -0.18)	0.02 (-0.04 to 0.08)
DET	-0.09 (-0.16 to -0.02)	0.01 (-0.09 to 0.10)	-0.26 (-0.32 to -0.20)	-0.23 (-0.29 to -0.18)	1	0.04 (-0.02 to 0.11)
LBE	0.16 (0.09 to 0.23)	0.11 (0.02 to 0.21)	0.06 (0.00 to 0.12)	0.02 (-0.04 to 0.08)	0.04 (-0.02 to 0.11)	1

Table 16: Estimates of association between IVF response variables from the correlated latent variable model. Posterior medians and 95% CIs.

10.5 Discussion

We have presented and compared several approaches for the analysis of multistage IVF data. All methods offer several advantages over those previously described, including the ability to incorporate mixed outcome types and responses defined at different levels of a multilevel data structure. The approaches are flexible enough to accommodate different combinations of response types and different covariates in the various submodels, according to the particular research question under consideration. The models can be fitted in freely available Bayesian software (Stan Development Team, 2017) without the need to write custom sampling algorithms.

The application to routinely collected clinical data highlighted a number of key differences between the approaches. The latent variable method can be used to examine relationships between covariates and stage-specific response variables. However, it is less useful for investigating the relationships between the responses, due to difficulties in interpreting the latent correlation coefficients and the fact that these cannot be adjusted for other response variables in the model. Both the outcome regression approach and the endogenous response model were preferable in this regard. Both provide easily interpretable regression coefficients and allow the causal structure of the sequence of responses to be represented. The validity of the estimates in the outcome regression approach rests upon an assumption of no unmeasured confounding however, which will always be implausible in practice. By contrast, the endogenous response model allows for the valid estimation of outcome-covariate effects by explicitly modelling the correlation between the error term and the response variable (Terza, 1998, Skrondal and Rabe-Hesketh, 2004). We have assumed a multivariate Gaussian distribution for the latent variables connecting the submodels. This is unverifiable in practice. Accordingly, the model should not be seen as a panacea for confounding. It might be possible to improve robustness in this regard using more flexible latent variable distributions, such as mixtures of Normals (Komarek et al., 2010) or copula-based methods (de Leon and Wu, 2011). We are aware that, in discussing the potential of these methods to quantify structural relationships in the IVF cycle, we have skirted the debate about whether or not it is meaningful to speak of causal effects of variables which are not directly modifiable (Greenland, 2017, Krieger and Davey Smith, 2016). The methods we present could be described using the language of causal mediation, so that instead of speaking of an effect of number of oocytes on downstream variables, for example, we could speak of the effects of predictors being mediated through the number of oocytes. A valid approach to mediation analysis in nonlinear models has been described by Pearl, (2011).

Identification of endogenous response models can be challenging. The inclusion of instrumental variables in some of the submodels is a common strategy to assist with identification. In our analysis, ‘attempt number’ acted as an instrumental variable in the number of oocytes and DET submodels and the binary variable ‘method of fertilization’ (either by mixing with sperm *in vitro* or by injecting the sperm directly into the oocytes) acted as an instrumental variable in the embryo quality submodels. We therefore assumed that attempt number affects the number of oocytes obtained and the decision to transfer two rather than one embryos (with previous failed attempts making it more likely both that higher doses of ovarian stimulation drugs will be used and that two rather than one embryo will be transferred) but does not influence the other response variables other than via these intermediaries. We also assumed that the method of fertilizing the oocytes influences the downstream outcomes solely through the quality of the resulting embryos. It is difficult to imagine how the method of fertilization could affect the cycle outcome by any other causal pathway. There could plausibly be unmeasured common causes of our instruments and downstream responses, which would undermine their validity as instruments. We note however that, since we handle endogenity through correlated latent variables in the model, validity of the instruments is not required. We anticipate that identification of endogenous response models is likely to be easier using larger datasets than that considered here, although as noted above fitting the models can be computationally expensive. It remains to identify a suitable reparametrization which may improve the speed of fitting the model, and to investigate the role of Bayesian prior regularization in improving efficiency (Gelman et al., 2014).

Parameter	Outcome regression model	Endogenous response model
Number of oocytes submodel		
Intercept	2.09 (2.07 to 2.12)	2.09 (2.07 to 2.12)
Age (SDs)	-0.18 (-0.21 to -0.15)	-0.18 (-0.21 to -0.15)
Partner Age (SDs)	0.04 (0.01 to 0.06)	0.03 (0.01 to 0.06)
Attempt number: 1st	0	0
2nd	0.06 (0.01 to 0.12)	0.07 (0.02 to 0.12)
3rd	0.17 (0.06 to 0.27)	0.15 (0.06 to 0.24)
4th or 5th	0.02 (-0.23 to 0.26)	0.14 (-0.10 to 0.37)
Fertilization rate submodel		
Intercept	-1.04 (-1.07 to -1.01)	-0.96 (-0.99 to -0.93)
Age (SDs)	0.06 (0.03 to 0.10)	0.07 (0.04 to 0.10)
Partner Age (SDs)	-0.02 (-0.05 to 0.02)	-0.02 (-0.05 to 0.01)
Embryo evenness submodel		
Intercepts (log odds of <=k): k=1	-4.33 (-4.47 to -4.18)	-4.33 (-4.49 to -4.19)
K=2	-1.37 (-1.47 to -1.28)	-1.37 (-1.47 to -1.27)
K=3	1.34 (1.24 to 1.43)	1.35 (1.25 to 1.45)
Age (SDs)	0.02 (-0.04 to 0.09)	0.02 (-0.09 to 0.13)
Partner Age (SDs)	0.04 (-0.02 to 0.11)	0.04 (-0.02 to 0.11)
Sperm injected into egg	-0.26 (-0.38 to -0.14)	-0.26 (-0.38 to -0.15)
Number of oocytes	0.09 (0.01 to 0.17)	0.24 (-0.10 to 0.60)
Fertilisation rate	-0.16 (-0.23 to -0.09)	-0.45 (-0.66 to -0.21)
Embryo fragmentation submodel		
Intercepts (log odds of <=k): k=1	-5.07 (-5.26 to -4.88)	-5.07 (-5.25 to -4.88)
K=2	-2.41 (-2.56 to -2.27)	-2.40 (-2.54 to -2.26)
K=3	-0.30 (-0.43 to -0.16)	-0.28 (-0.41 to -0.15)
Age (SDs)	-0.12 (-0.21 to -0.03)	-0.22 (-0.37 to -0.07)
Partner Age (SDs)	0.07 (-0.02 to 0.16)	0.08 (-0.02 to 0.17)
Sperm injected into egg	-0.32 (-0.48 to -0.16)	-0.32 (-0.48 to -0.15)
Number of oocytes	0.22 (0.09 to 0.34)	-0.09 (-0.53 to 0.42)
Fertilisation rate	-0.30 (-0.41 to -0.20)	-0.57 (-0.90 to -0.20)
Double embryo transfer submodel		
Intercept	0.13 (0.07 to 0.19)	0.08 (0.02 to 0.14)
Age (SDs)	0.07 (0.00 to 0.13)	-0.03 (-0.11 to 0.05)
Partner Age (SDs)	-0.03 (-0.09 to 0.03)	-0.02 (-0.07 to 0.04)
Attempt No: 1st	0	0
2nd	0.25 (0.12 to 0.37)	0.26 (0.15 to 0.38)
3rd	0.47 (0.23 to 0.70)	0.53 (0.31 to 0.76)
4th or 5th	0.63 (0.08 to 1.22)	0.68 (0.15 to 1.23)
Number of oocytes	-0.06 (-0.13 to 0.02)	-0.25 (-0.48 to 0.01)
Fertilization rate	-0.02 (-0.09 to 0.05)	-0.37 (-0.53 to -0.16)
Embryo evenness	-0.14 (-0.20 to -0.08)	-0.09 (-0.15 to -0.02)
Embryo fragmentation	-0.12 (-0.18 to -0.06)	-0.04 (-0.12 to 0.03)
Live birth event submodel		
Intercept	-0.55 (-0.64 to -0.47)	-0.38 (-0.73 to -0.06)
Age (SDs)	-0.06 (-0.12 to -0.00)	-0.12 (-0.20 to -0.04)
Partner Age (SDs)	-0.05 (-0.11 to 0.02)	-0.03 (-0.09 to 0.03)
Number of oocytes	0.04 (-0.03 to 0.13)	-0.11 (-0.35 to 0.14)
Fertilization rate	0.14 (0.08 to 0.21)	-0.17 (-0.35 to 0.03)
Embryo evenness	0.07 (0.01 to 0.13)	0.03 (-0.04 to 0.10)
Embryo fragmentation	0.04 (-0.24 to 0.10)	0.04 (-0.04 to 0.11)
DET	0.11 (0.00 to 0.21)	-0.15 (-0.63 to 0.44)

Table 17: Regression coefficients from outcome regression and endogenous response models for the IVF cycle. Posterior medians and 95% CIs.

Although the latent variable model was not useful for the purpose of investigating relationships between responses, models of this sort can be usefully employed for the purpose of making multivariate predictions (McCulloch, 2008). Using the posterior draws from the latent variable model fitted in section 10.4.3, we predicted the IVF cycle outcomes for a new cohort of patients with the same characteristics as those in our sample. We adopted a sequential approach whereby we predicted the number of oocytes obtained after stimulation for each patient for each draw from the posterior, before predicting the fertilization rate (and consequently, the number of embryos obtained) in those who were predicted to have any oocytes available. We then predicted the embryo quality for each embryo predicted to arise from the fertilization procedure, and so on. This approach allows us to predict the responses of a cohort of patients (or indeed, of an individual) as they pass through the stages of the IVF cycle, while incorporating the dependency between stage-specific responses. This is not a feature of existing prediction models (eg: Dhillon et al., 2016, Nelson and Lawlor, 2011), but may be useful to the clinician whenever there is interest not only in the overall outcome of treatment but also in ensuring that this is achieved in a safe manner. For example, large egg yields following ovarian stimulation are associated with increased risk of ovarian hyperstimulation syndrome (Steward et al., 2014), low birthweight and preterm birth (Sunkara et al., 2015). Consequently, a target of ovarian stimulation is to obtain a yield of oocytes which is neither too low to limit the overall likelihood of success, nor too high to represent a risk to the patient or offspring (La Marca and Sunkara 2014, La Marca et al., 2012). We can use the latent variable approach to predict the probability of a patient obtaining a safe yield of oocytes under a given treatment (eg: fewer than 15) and going on to have a live birth. Without conditioning on any patient or treatment covariates, we calculate this as 23%, with a 95% prediction interval of 21 to 25%. It remains to establish whether there is any advantage offered by including outcome-covariates in the prediction setting.

While all of the models presented here can accommodate embryo-level response variables, relationships between these and other outcomes are estimated using the mean value (Dunson et al., 2003). An undesirable consequence of this is the implicit assumption that the relationship between the evenness and fragmentation of an embryo is the same as the relationship between the evenness of an embryo and the fragmentation of another from the same patient (Gueorguieva, 2001). This could be relaxed by using latent representations of the embryo grading submodels and allowing the embryo-level residual terms to be correlated (McCulloch, 2008, Gueorguieva and Sanacora, 2006). A related concern is the fact the models do not allow embryo-level responses to be included as covariates in the DET and LBE submodels without averaging the values over a patient's embryos. The estimation of the effects of embryo characteristics on birth outcomes is complicated by the fact that if two embryos are transferred and only one implants, it is not known which of the two was successful. This partial observability problem motivates the use of embryo-uterine models which have been described from both Bayesian (Dukic and Hogan, 2002) and Likelihood (Roberts, 2007) viewpoints. It remains to incorporate the embryo-uterine approach in the joint modelling approaches described here. We also note that the mean value might not be the best summary measure to use for the purpose of including the embryo gradings as covariates in the DET and LBE submodels, since the best one or two are selected for transfer. An alternative measure capturing the highest available grades might be more appropriate future applications of the methods. Alternatively, we could include the quality of the transferred embryos as additional embryo-level response variables in the model. More generally, we note the fact that we included only a small number of covariates as a limitation of our analysis. We anticipate that differences between the outcome regression and endogenous response approaches will reduce if further covariates are available to control for confounding. This is a topic for future research.

In these examples, we do not differentiate between twin and singleton births. The difference is clinically important, since twin pregnancies represent increased risk to the mother and infants. While we do not make this distinction here, any of the approaches we describe could be extended to accommodate a twin vs singleton submodel, fitted conditional on birth. Our live birth submodel also does not distinguish between failure due to transferred embryos not implanting in the uterine wall and failure due to implanted embryos not being sustained to term (ie: miscarriage). This may be an important distinction for some research questions. Separate submodels for embryo implantation and birth (conditional on implantation) could be included to this end.

All of the approaches presented here assume that any drop out from the cycle is MAR. This might be plausible, since drop out is usually the result of poor response or outright failure at one stage, preventing further progression. These response variables are included in the models. If however, the MAR assumption is deemed not to be appropriate, we could jointly model the drop out process by defining a sequential probit submodel (Albert and Chib, 2001) corresponding to transitions through the stages, and allowing this to be correlated with the stage-specific responses (Steele et al., 2009). An alternative strategy would be to employ a selection modelling approach (Heckman, 1976, 1978, Diggle and Kenward, 1994) where the probability of dropout at a given stage is related to the coincident (possibly unobserved) response variables by way of inclusion as covariates and/or correlated latent variables. Selection models are difficult to implement in Stan, which does not support discrete parameters, thereby precluding explicit modelling of missing egg counts or ordinal gradings. More generally, we might question whether *missing-not-at-random* (MNAR) methods are suitable in the present context. Given that downstream responses are defined conditional on upstream success, this may be construed as an example of the phenomenon known as *truncation-by-death* (Zhang and Rubin, 2003, Rubin, 2006). McConnell and colleagues (2008) note that MNAR methods implicitly assume an underlying value for missing outcomes, and discuss principal stratification approaches as an alternative. The applicability of principal stratification methods to complex multistage IVF data warrants consideration in future research.

10.6 Recommendations

When analysing multistage IVF data, the appropriate analytic method will depend on the exact research question under consideration. If interest is in estimating the effects of treatment and patient characteristics on outcomes, as well as the structural relationships between the responses at each stage, we recommend the use of the endogenous response model. Identification of the model is likely to require relatively large, detailed datasets, and researchers should be realistic about the scope to answer mechanistic research questions where this resource is not available. Questions of this sort may be tackled using the outcome regression approach, but researchers should be aware that this involves the strong assumption that confounding has been adequately dealt with by measured covariates. In our simple example, we arrived at substantively different conclusions regarding the effects of procedural response variables on downstream outcomes in the outcome regression approach compared to the endogeneity approach, which allows for the correlation between procedural responses and unmeasured predictive factors. We would urge caution when interpreting the endogenous response model however, since inevitable misspecification of the latent variable distribution means that residual confounding will not be eliminated. Researchers should still attempt to reduce confounding through the inclusion of known variables as far as possible. Estimates corresponding to other (exogenous) covariates were similar between models. Where interest lies in making predictions about the patient journey through the stages of the IVF cycle, the relatively simple latent variable approach offers a framework to do this while taking the dependency between the stages into consideration. These approaches assume that drop out is MAR. We are unable to make a definitive recommendation regarding the most appropriate approach to modelling drop out at present, other than to state that MNAR methods assume that there is an underlying value for each missing response. This may not be appropriate where responses are strictly undefined. Finally, given the complexity of IVF, we note that any meritorious analysis will require substantial input from clinician and clinical scientist collaborators.

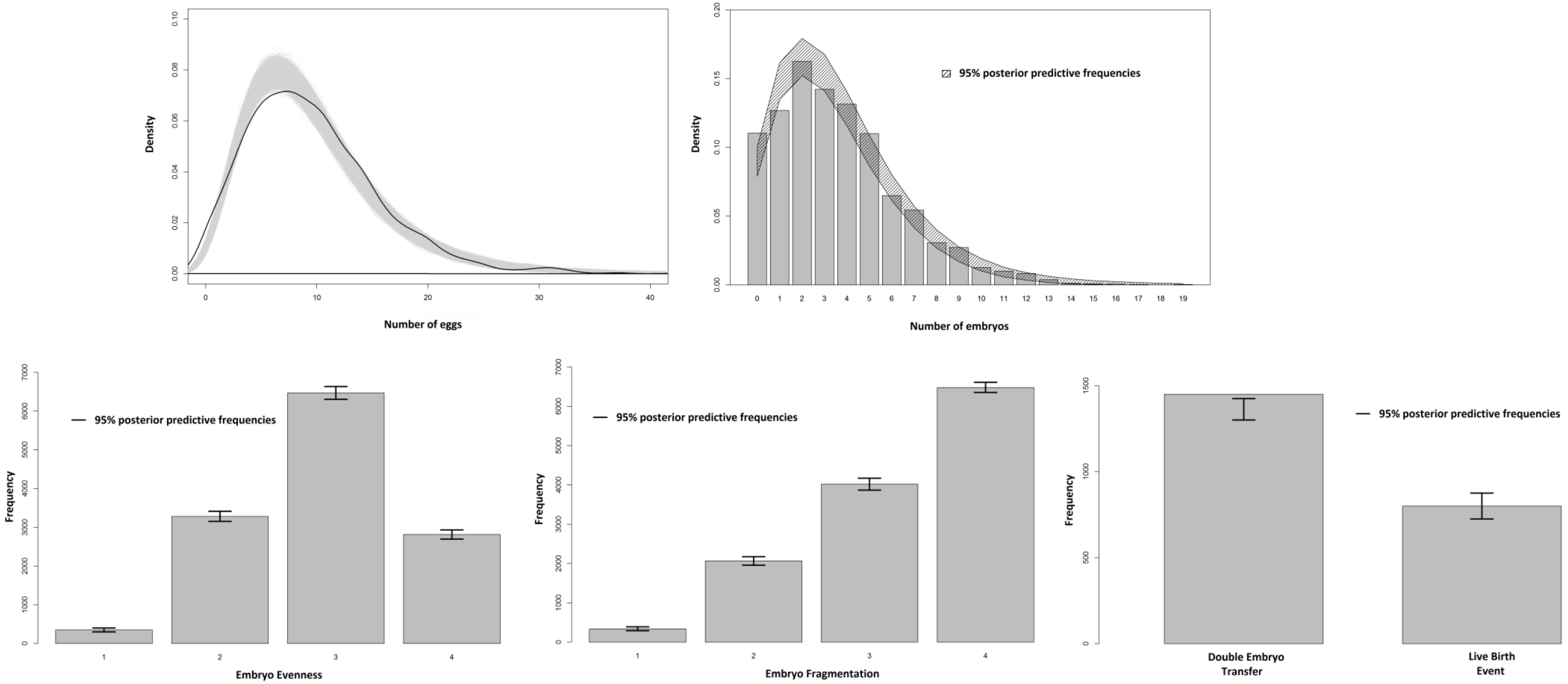


Figure 39: Observed response distribution against simulated datasets drawn from the posterior predictive distribution. 'Error bars' on barplots are 95% intervals based on the simulated proportions

10.7 **References for Chapter 10.**

Albert, J.H. and Chib, S. Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass* 1993: 88(422), pp.669-679.

Albert, J. H. and Chib, S. Sequential ordinal modeling with applications to survival data. *Biometrics* 2001: 57(3), pp. 829-836.

Blalock, H. M. Evaluating the Relative Importance of Variables. *American Sociological Review* 1961, 26(6), pp. 866-874.

De Leon, A. R. and Chough, K. C. Analysis of mixed data: methods & applications, 2013. CRC Press.

de Leon, A. R. and Wu, B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Stat Med* 2011: 30(2), pp. 175-85.

Dhillon, R. K., McLernon, D. J., Smith, P. P., Fishel, S., Dowell, K., Deeks, J. J., Bhattacharya, S. and Coomarasamy, A. Predicting the chance of live birth for women undergoing IVF: a novel pretreatment counselling tool. *Hum Reprod* 2016: 31(1), pp. 84-92.

Diggle, P., Farewell, D. and Henderson, R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *J R Statist Soc Ser C Appl Stat* 2007: 56, pp. 499-529.

Diggle, P. and Kenward, M. G. Informative Drop-out in Longitudinal Data-Analysis. *J R Statist Soc Ser C Appl Stat* 1994: 43(1), pp. 49-93.

Dukic, V. and Hogan, J. W. A hierarchical Bayesian approach to modeling embryo implantation following in vitro fertilization. *Biostatistics* 2002: 3(3), pp. 361-377.

Dunson, D. B. Bayesian latent variable models for clustered mixed outcomes. *J R Stat Soc Series B Stat Methodol* 2000: 62, pp. 355-366.

Dunson, D. B., Chen, Z. and Harry, J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* 2003: 59(3), pp. 521-530.

Dunson, D. B. and Herring, A. H. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* 2005: 6(1), pp. 11-25.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. Bayesian data analysis, 2014. CRC press.

Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat Sci* 1992: pp. 457-472.

Goldstein, H. Multivariate Multilevel Data. in Multilevel Statistical Models, 3rd Edition, 2003. pp. 139-146. Arnold.

Goldstein, H. 3-level Models and more Complex Hierarchical Structures. in Multilevel Statistical Models, 4th Edition, 2011. pp. 73-110. John Wiley & Sons.

Goldstein, H., Carpenter, J., Kenward, M. G. and Levin, K. A. Multilevel models with multivariate mixed response types. *Stat Model* 2009: 9(3), pp. 173-197.

Gould, A. L., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S. and Bois, F. Y. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Stat Med* 2015: 34(14), pp. 2181-2195.

Greenland, S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *Eur J Epidemiol* 2017.

Greenland, S., Maclure, M., Schlesselman, J. J., Poole, C. and Morgenstern, H. Standardized Regression-Coefficients - a Further Critique and Review of Some Alternatives. *Epidemiology* 1991: 2(5), pp. 387-392.

Gueorguieva, R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat Model* 2001: 1(3), pp. 177-193.

Gueorguieva, R. V. and Agresti, A. A correlated probit model for joint modeling of clustered binary and continuous responses. *J Am Stat Assoc* 2001:, 96(455), pp. 1102-1112.

Gueorguieva, R. V. and Sanacora, G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat Med*. 2006: 25(8), pp. 1307-22.

Heckman, J. J. Common Structure of Statistical-Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. Annals of Economic and Social Measurement 1976: 5(4), pp. 475-492.

Heckman, J. J. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* 1978: 46(4), pp. 931-959.

Heijnen, E., Macklon, N. S. and Fauser, B. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod* 2004: 19(9), pp. 1936-1938.

Komarek, A., Hansen, B. E., Kuiper, E. M., van Buuren, H. R. and Lesaffre, E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Stat Med* 2010: 29(30), pp. 3267-83.

Krieger, N. and Davey Smith, G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol* 2016.

La Marca, A., Papaleo, E., Grisendi, V., Argento, C., Giulini, S. and Volpe, A. Development of a nomogram based on markers of ovarian reserve for the individualisation of the follicle-stimulating hormone starting dose in in vitro fertilisation cycles. *BJOG* 2012: 119(10), pp. 1171-1179.

La Marca, A. and Sunkara, S. K. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Hum Reprod Update* 2014: 20(1), pp. 124-140.

Legro, R. S., Wu, X. K., Barnhart, K. T., Farquhar, C., Fauser, B. C. J. M., Mol, B., Conference, H. C. and Comm, S. Improving the Reporting of Clinical Trials of Infertility Treatments (IMPRINT): modifying the CONSORT statement. *Hum Reprod* 2014: 29(10), pp. 2075-2082.

Maity, A., Williams, P. L., Ryan, L., Missmer, S. A., Coull, B. A. and Hauser, R. Analysis of in vitro fertilization data with multiple outcomes using discrete time-to-event analysis. *Stat Med*. 2014: 33(10), pp. 1738-1749.

McConnell, S., Stuart, E. A. and Devaney, B. The truncation-by-death problem - What to do in an experimental evaluation when the outcome is not always defined. *Evaluation Rev* 2008: 32(2), pp. 157-186.

McCulloch, C. Joint modelling of mixed outcome types using latent variables. *Stat Methods Med Res* 2008: 17(1), pp. 53-73.

Min, J. K., Breheny, S. A., MacLachlan, V. and Healy, D. L. What is the most relevant standard of success in assisted reproduction? The singleton, term gestation, live birth rate per cycle initiated: the BESST endpoint for assisted reproduction. *Hum Reprod* 2004: 19(1), pp. 3-7.

Nelson, S. M. and Lawlor, D. A. Predicting Live Birth, Preterm Delivery, and Low Birth Weight in Infants Born from In Vitro Fertilisation: A Prospective Study of 144,018 Treatment Cycles. *PLoS Med* 2011: 8(1).

Pearl J. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In Causality: Statistical Perspectives and Applications (eds C. Berzuini, P. Dawid, L. Bernardinelli), 2011: ch. 12,pp.151-175. Hoboken, John WIley & Sons.

Pearl, J. Direct and indirect effects. Proceedings of the seventeenth conference on uncertainty in artificial intelligence 2001: pp. 411-20. Seattle, Washington.

Penman, R., Heller, G. and Tyler, J. Modelling IVF Data using an Extended Continuation Ratio Random Effects Model. in Proceedings of the 22nd International Workshop on Statistical Modelling 2007: pp. 482-485. Barcelona.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2014. Vienna, Austria.

Roberts, S., Hirst, W., Brison, D., Vail, A. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. *Hum Reprod* 2010: 25(11), pp. 2792-2802.

Roberts, S. A. Models for assisted conception data with embryo-specific covariates. *Stat Med* 2007: 26(1), pp. 156-170.

Rubin, D. B. Inference and Missing Data. *Biometrika* 1976: 63(3), pp. 581-590.

Rubin, D. B. Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Stat Sci* 2006: 21(3), pp. 299-309.

Skrondal, A. and Rabe-Hesketh, S. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models, 2004. CRC Press.

Stan Development Team. RStan: the R interface to Stan, Version 2.5.0, 2014.

Stan Development Team. Stan Modeling Language: User's Guide and Reference Manual, 2017[online], available: http://mc-stan.org/users/documentation/

Steele, F., Sigle-Rushton, W. and Kravdal, O. Consequences of Family Disruption on Children's Educational Outcomes in Norway. *Demography* 2009: 46(3), pp. 553-574.

Sterne, J. A., Hernan, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hrobjartsson, A., Kirkham, J., Juni, P., Loke, Y. K., Pigott, T. D., Ramsay, C. R., Regidor, D., Rothstein, H. R., Sandhu, L., Santaguida, P. L., Schunemann, H. J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J. C., Waddington, H., Waters, E., Wells, G. A., Whiting, P. F. and Higgins, J. P. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *Bmj* 2016: 355, pp. i4919.

Steward, R. G., Lan, L., Shah, A. A., Yeh, J. S., Price, T. M., Goldfarb, J. M. and Muasher, S. J. Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: an analysis of 256,381 in vitro fertilization cycles. *Fertil Steril* 2014: 101(4), pp. 967-973.

Sunkara, S. K., La Marca, A., Seed, P. T. and Khalaf, Y. Increased risk of preterm birth and low birthweight with very high number of oocytes following IVF: an analysis of 65 868 singleton live birth outcomes. *Hum Reprod* 2015: 30(6), pp. 1473-1480.

Terza, J. V. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *J Econom* 1998: 84(1), pp. 129-154.

Terza, J. V. Parametric Nonlinear Regression with Endogenous Switching. *Econom Rev* 2009: 28(6), pp. 555-580.

Tsiatis, A. A. and Davidian, M. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* 2004: 14(3), pp. 809-834.

Westreich, D. and Greenland, S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol* 2013: 177(4), pp. 292-8.

Wilkinson, J., Roberts, S. A., Showell, M., Brison, D. R. and Vail, A. (2016) No common denominator: a review of outcome measures in IVF RCTs. *Hum Reprod* 2016: 31(12), pp. 2714-2722.

Xie, Y. Endogenous Switching Regression Models. in Into Adulthood: A Study of the Effects of Head Start 2000: pp. 169-178.

Zhang, J. N. L. and Rubin, D. B. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J Educ Behav Stat* 2003: 28(4), pp. 353-368.

10.8 Supplementary material for Chapter 10.

	Number of oocytes	Fertilization rate	Embryo evenness	Embryo fragmentation	DET	LBE
Number of oocytes	1	-0.58 (-0.64 to -0.51)	0.00 (-0.35 to 0.29)	0.29 (-0.02 to 0.51)	0.42 (0.17 to 0.62)	0.36 (0.09 to 0.59)
Fertilization rate	-0.58 (-0.64 to -0.51)	1	0.34 (-0.01 to 0.58)	0.06 (-0.23 to 0.32)	0.27 (-0.04 to 0.55)	0.23 (-0.05 to 0.50)
Embryo evenness	0.00 (-0.35 to 0.29)	0.34 (-0.01 to 0.58)	1	0.84 (0.76 to 0.88)	-0.06 (-0.22 to 0.07)	0.08 (-0.06 to 0.21)
Embryo fragmentation	0.29 (-0.02 to 0.51)	0.06 (-0.23 to 0.32)	0.84 (0.76 to 0.88)	1	-0.02 (-0.15 to 0.12)	0.09 (-0.05 to 0.22)
DET	0.42 (0.17 to 0.62)	0.27 (-0.04 to 0.55)	-0.06 (-0.22 to 0.07)	-0.02 (-0.15 to 0.12)	1	0.32 (-0.05 to 0.60)
LBE	0.36 (0.09 to 0.59)	0.23 (-0.05 to 0.50)	0.08 (-0.06 to 0.21)	0.09 (-0.05 to 0.22)	0.32 (-0.05 to 0.60)	1

S Table 21: Latent correlation matrix derived from the endogeneity model of the IVF cycle. All submodels adjusted for age, partner age. Number of oocytes and DET submodels additionally adjusted for attempt number. Embryo quality submodels additionally adjusted for method of fertilization (mixing eggs with sperm in vitro or injecting sperm directly into the egg). Posterior medians and 95% CIs.

Chapter 11. Does ovarian stimulation with gonadotropins affect uterine receptivity? A multistage modelling study.

Journal article 6

Authors Jack Wilkinson, Daniel R Brison, Cheryl Fitzgerald, Andy Vail, Stephen A Roberts

Status Not yet submitted

Contribution statement JW devised the study, prepared the dataset, conducted and interpreted statistical analysis and co-authored the manuscript. SR and AV devised the study, contributed to the interpretation of statistical analysis and co-authored the manuscript. DRB and CF contributed to the interpretation of statistical analysis and co-authored the manuscript.

Preamble In this paper we use the multistage modelling methodology developed in the thesis to investigate a clinical research question: does ovarian stimulation affect the uterine environment, making it less likely that embryos will implant and be carried to term? We add a submodel corresponding to the total dose of gonadotropins received by the patient, and use two submodels to represent the outcomes of the transfer process; embryo implantation and live birth event (given embryo implantation). We felt it was important to make the distinction between implantation (only) and live birth for the present study, since our interest is in the specific effects of stimulation on the patient. This research question is difficult to answer, because ovarian stimulation influences the clinical outcome by expanding the available embryo pool in addition to any physiological effects on the uterus. Separating these two causal pathways is difficult. Standard outcome regression (for example, a logistic regression of embryo implantation with dose, egg yield, embryo quantity and quality and other predictors included as covariates) would not be sufficient, because dose and the other response variables are endogenous.

In this instance, the variables accounting for a large portion of the unmeasured confounding are known, but unmeasured; a patient's ovarian reserve is used to determine the starting dose, and ongoing monitoring results in dose switching. We employ instrumental variables and correlated latent variables in an attempt to soak up residual confounding (see the illustrative simulation study in 7.7).

Outputs and Impact of the research This work has not yet been submitted to a journal, although an abstract has been submitted to the Fertility 2018 conference.

11.1 Abstract

Study question

Does controlled ovarian stimulation (COS) with gonadotropins affect the uterine environment, making it less likely that transferred embryos will implant?

Summary answer

After controlling for confounding and the influence of COS on oocytes and embryos, we found that higher gonadotropin doses resulted in a reduced chance that embryos would implant and be carried to term.

What is known already

It has been suggested that COS adversely affects endometrial angiogenesis, making it less likely that embryos will successfully implant and develop in utero. This motivates the idea of freezing all embryos, and delaying transfer until the woman has recovered from stimulation. Previous studies have suffered from methodological limitations, however.

Study design, size, duration

Multistage modelling of a routine ART database. Analysis allowed for COS to influence the final outcome both by increasing the pool of available embryos and by having physiological effects on the uterine environment, and for the separate estimation of these effects. Analysis also allowed for unmeasured confounding between gonadotropin dose and patient responses. 2968 cycles in 2457 patients in the three-year time period January 2013-December 2015 were included.

Participants/materials, setting, methods

Women aged 21-43 years commencing COS for IVF or ICSI at the Department of Reproductive Medicine, St Mary's Hospital, Manchester, England.

Main results and the role of chance

Total gonadotropin doses ranged between 224 and 7650IU, with a median of 2250IU. After controlling for both measured and unmeasured confounding due to patient characteristics, a dose-response relationship with embryo implantation was evident, with increasing total gonadotropin dose resulting in reduced likelihood of embryo implantation across the dose range. The likelihood that an implanted embryo would be sustained to birth was reduced up to a dose of around 1900IU, before levelling off. The dose effect appeared to be attenuated in blastocyst as opposed to cleavage stage transfers.

Limitations, reasons for caution

While we adjusted both for measured confounding variables and for unmeasured confounding arising due to the fact that dose is selected on the basis of patient characteristics, there could still be some confounding due to measurement error in the model.

Wider implications of the findings

The present study supports the theory that COS has deleterious consequences for the uterine environment and for embryo implantation and development. While uncertainty remains as to the effectiveness of elective cryopreservation cycles, blastocyst transfers may mitigate these effects.

Study funding/ competing interests

JW is funded by a Doctoral Research Fellowship from the National Institute for Health Research (DRF-2014-07-050) supervised by AV and SAR. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. JW and AV are statistical editors of the Cochrane Gynaecology and Fertility Group. AV has received fees from the Human Fertilisation and Embryology Authority, outside the submitted work. SR is a

statistical editor for Hum Reprod. JW also declares that publishing peer-reviewed articles benefits his career.

11.2 Introduction

In controlled ovarian stimulation (COS), the ovaries are stimulated with gonadotropins. The aim is to furnish the embryologist with a sufficient number of oocytes to produce a robust pool of embryos. She can then select the most suitable candidate amongst these for transfer. However, while COS improves the chances of obtaining high quality embryos for fresh transfer and freezing, it is also associated with deleterious consequences, including risk of ovarian hyperstimulation syndrome (OHSS) (Steward, et al., 2014). Additionally, it has been suggested that COS might negatively impact upon uterine receptivity, making it less likely that transferred embryos will implant (Maheshwari and Bhattacharya, 2013). The evidence in support of this idea is limited however, since the task of distinguishing adverse effects on the uterine environment from the benefits of an expanded stock of embryos presents a considerable statistical challenge. Several observational studies have indirectly supported the hypothesis by suggesting increased risk of obstetric haemorrhage (Healy, et al., 2010), and of perinatal morbidity (Kalra, et al., 2011), in fresh compared to frozen transfer cycles. This has been attributed to the post-COS recovery period afforded to the woman by the latter. However, although the authors of these studies attempted to adjust for confounding, the multivariable regression methods they employed are not capable of untangling the causal web that includes COS, oocyte yield, embryo selection and cryopreservation amongst its strands (Blalock, 1961, Westreich and Greenland, 2013). Confounding is not the only impediment. Strong selection effects also obfuscate the interpretation of these studies. Patients who experience adverse outcomes after a fresh transfer might be less likely to undergo subsequent treatment. Susceptible patients will therefore be underrepresented in frozen cycles, wherever these are not restricted to initial elective treatments. In addition, patients who have had all embryos frozen in order to mitigate OHSS will not be present in fresh cycle data. Pregnancy rates in this group are generally high (Raziel, et al., 2009). Similarly, an extreme form of selection bias casts doubt on several recent studies comparing obstetric and perinatal outcomes in frozen compared to fresh transfer cycles, including a recent highly cited meta-analysis (Maheshwari, et al., 2012), and a retrospective analysis of births from two RCTs (Shapiro, et al., 2016). This arises wherever miscarriages are excluded from analysis, since deleterious effects of treatment on gestation are likely to manifest in unsuccessful pregnancies. This is the same fallacy that led to erroneous conclusions in studies of folate supplementation for the prevention of neural tube defects (Hernan, et al., 2002). Differential expression of genes involved in endometrial receptivity after COS has been suggested in an exploratory study (Haouzi, et al., 2009), although no confirmatory study appears to have been conducted to affirm these results.

Consequently, much of the evidence suggesting that COS adversely affects the uterine environment in humans is indirect and methodologically limited. Nonetheless, there is understandably considerable interest in the effectiveness of frozen cycles (Maheshwari and Bhattacharya, 2013, Weinerman and Mainigi, 2014); a recent Cochrane Review found insufficient evidence to reach a conclusion in this regard from four completed RCTs, but identified 12 ongoing trials (Wong, et al., 2017). While it is unclear what this body of research will eventually show, it remains in the interim to identify the underlying causal mechanisms by which COS influences the IVF outcome, using appropriate statistical methods. Mechanistic understanding of complex interventions is necessary for explaining the success and failure of tested treatments, as well as for the design and testing of new treatments (Emsley, et al., 2010).

Joint modelling techniques have long been used to estimate causal effects from observational data in the presence of selection biases and unmeasured confounding (eg: (Heckman, 1976, Heckman, 1978, Skrondal and Rabe-Hesketh, 2004, Terza, 1998). Sufficient computational power is now readily available to permit the adaptation and application of these methods to large datasets and complex multilevel data structures (Dunson, 2000, Dunson, et al., 2003, Goldstein, et al., 2009, Steele and Washbrook, 2013). This includes complex multistage treatments such as IVF. We developed a model in this framework in order to investigate how COS impacts downstream events, including oocyte yield, fertilisation, embryo morphology, implantation and development in utero. This approach allows us to distinguish 'direct' effects of COS from 'indirect' effects influencing the transfer outcome by way of expansion of the embryo pool and the quality of the embryos selected for transfer (Skrondal and Rabe-Hesketh, 2004). It also estimates treatment effects while allowing for unmeasured confounding (Skrondal and Rabe-Hesketh, 2004, Terza, 1998). We used the model to assess COS effects in a large routine ART database. We present the findings from our analysis and discuss the implications.

11.3 Materials and methods

11.3.1. Population

We included women aged over 18 years commencing COS for IVF at Department of Reproductive Medicine, St Mary's Hospital, Manchester, England. The three-year time period January 2013-December 2015 was considered. Donor and banking cycles were excluded. Small numbers of cycles were excluded for no sperm (5 cycles), for loss to follow up (26), for missing embryo data (53) and missing starting dose (3).

Cause of subfertility was established by an initial investigation. Male factor subfertility was established if the partner had azoospermia, surgical sperm extraction, or severe oligiospermia, or if sperm counts or other parameters were abnormal. Endometriosis was diagnosed where there was a history of endometriosis confirmed by laparoscopy. Endometrioma diagnosis was based on an ultrasound scan. If no cause was identified, we considered the reason for subfertility to be unexplained.

11.3.2. ART Protocols

Patients were allocated to a starting dose of gonadotropins (either hMG, (Menopur, Ferring Pharmaceuticals Ltd, UK) or rFSH (Gonal F, Merck Serono Ltd)) on the basis of anti-mullerian hormone levels, antral follicle count, and previous response to stimulation in either a GnRH Long Agonist or GnRH Antagonist protocol, depending on patient preference and clinician judgement based on individual clinical features. In the Long Agonist protocol, 0.25mg of Buserelin (Supercur, Sanofi Aventis Ltd., Surrey, UK) was administered daily starting from the mid-luteal phase (day 21 in a 28-day cycle) and continuing until the day of trigger. In the Antagonist protocol, 0.25mg of Cetrotide (Cetrorelix, Merck Serono Ltd, Middlesex, UK) was administered from day 5 until the day of trigger. The cycle was monitored by serum oestradiol (E2) and transvaginal ultrasound scan beginning on stimulation day 6 in cycles at high risk of excessive response and on day

8 otherwise. The daily FSH dose could be altered from day 6 on the basis of ultrasound results at the discretion of the clinician. The criteria for triggering of oocyte maturation were 3 or more follicles > 17mm on ultrasound and E2 < 15,000 pmol/L. In Long Agonist cycles, triggering was performed using 5,000 IU of HCG (Pregnyl, Organon Laboratories Ltd., Cambridge, UK), or 10,000 IU in the event that 4 or fewer eggs were expected. The same triggering protocol was used in Antagonist cycles, unless an excessive response was anticipated, in which case an Agonist trigger was used.

Oocyte pickup (OPU) took place around 34-36 hours following HCG trigger, guided by ultrasound. Aspirated oocytes were immediately identified and counted by an embryologist following the procedure. Embryos were graded according to the British Fertility Society and Association of Clinical Embryologist joint guidelines for elective single embryo transfer (Harbottle, et al., 2015). We use the day 2 gradings (approx. 66 hrs after insemination) for analysis.

11.3.3. **Statistical Analysis**

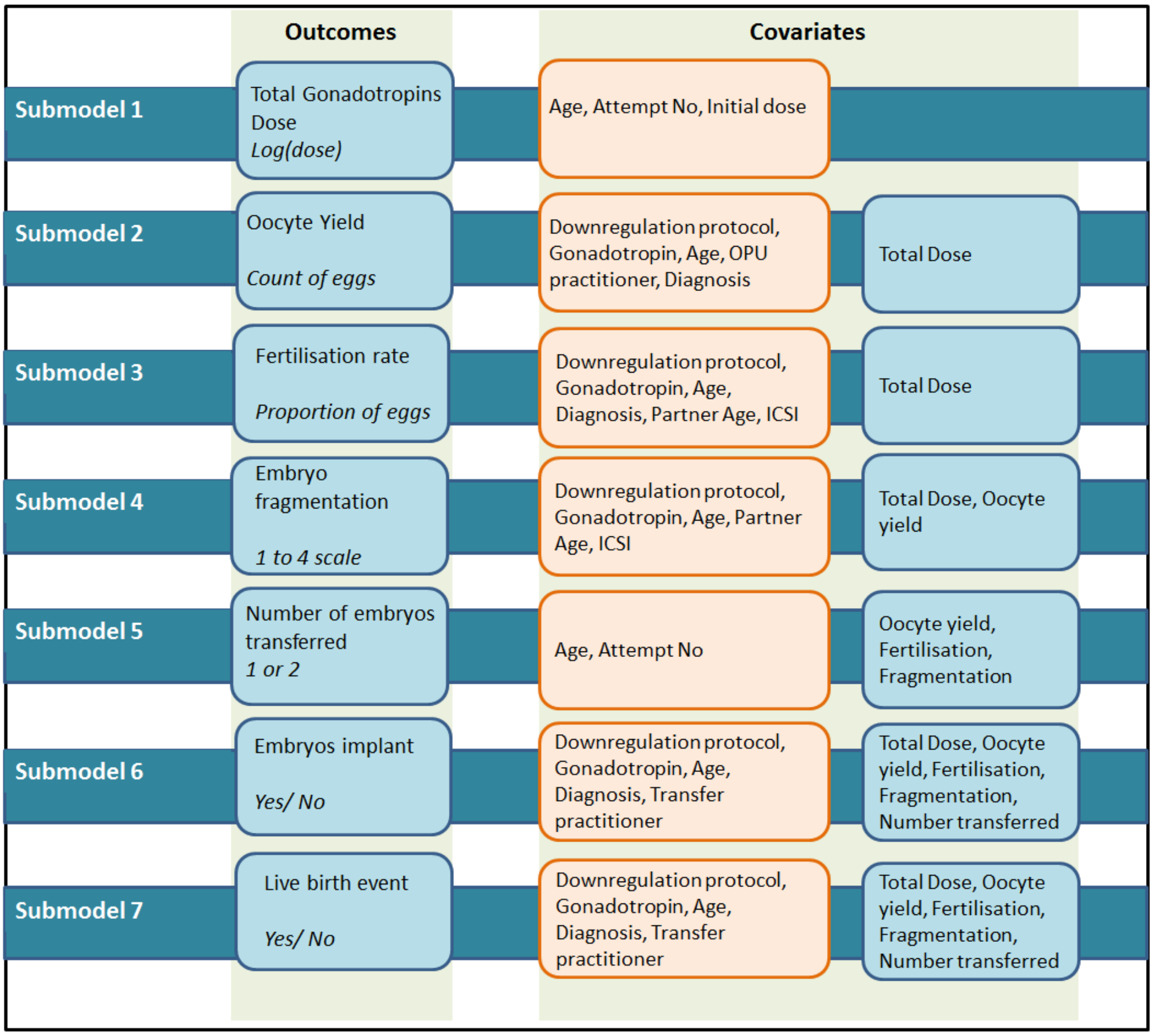
Our model comprises several regression submodels corresponding to the sequence of treatments and responses arising in the fresh IVF cycle. The response variables in the model are total dose of gonadotropins administered, number of oocytes obtained in OPU, number of cleaved embryos obtained from fertilisation, embryo-level morphology grading (degree of fragmentation), the number of embryos transferred, embryo implantation (whether or not at least one implanted) and whether or not implanted embryos were carried to term (live birth event). The treatment variables total dose and number of embryos transferred are included as response variables because both are determined by patient characteristics and, in the case of the latter, additionally by outcome of stimulation, fertilisation and culture. In particular, dose is increased to compensate for anticipated poor ovarian response. This induces confounding by indication (Walker, 1996) between dose and patient outcomes, which must be corrected for. By including dose as a response variable, we can model the correlation introduced by the confounding. This allows us to adjust our estimated dose effects for confounders, even though some of these are not measured (Heckman, 1976, Heckman, 1978, Skrondal and Rabe-Hesketh, 2004).

The representations and regression types used for each response were: total dose (log transform, linear regression); number of oocytes (count variable, Poisson regression); fertilisation rate (count variable, Poisson regression with number of eggs as an offset); embryo fragmentation (ordinal 1 to 4 variable, cumulative logit regression); number of embryos transferred (1 or 2 embryos, probit regression); embryo implantation (any or none, probit regression); and live birth event (yes or no, probit regression). We included only fragmentation grade, and not evenness, as a measure of embryo quality because both were strongly correlated, meaning that it was not possible to statistically distinguish between these two in the model. Evenness has previously been suggested as having no additional predictive value after fragmentation has been taken into account (Stylianou, et al., 2012). We used the day 2 fragmentation grading in the analysis, since by day 3 some cleaved embryos are lost, with the potential to introduce truncation biases into the analysis.

We represented total dose using regression splines, based on a B-spline basis, using the splines package in R (R Core Team, 2014). We included knots at the variable quintiles (Harrell Jr, 2015). This allows for nonlinear effects of dose on the downstream response variables.

As noted above, we allowed for unmeasured confounding between the response variables in the model, by explicitly representing this as unexplained covariation (Terza, 1998). We also minimised unmeasured confounding as far as possible by including known predictive variables in the appropriate submodels. Figure 40 gives full details of the covariates included in each part (submodel) of the model. S Table 22 to S Table 30 give additional information regarding their representational forms. Each submodel is fitted to those patients who did not drop out at an earlier stage of treatment (so that patients with no oocytes are not included in the submodels of fertilisation, embryo morphology, or in any of the submodels pertaining to embryo transfer, for example). This amounts to an assumption that missing data are explicable by the response and covariate data included in our model (Rubin, 1976). We use Bayesian regularisation to fit the model (Gelman, et al., 2014). Briefly, this allows complex models to be fitted by excluding a priori highly implausible parameter values from consideration. Even a large dose increase will not result in a 100-fold increase in the number of eggs obtained, for example. 11.7.3 gives the mathematical representation of the model and S Figure 3 shows a path diagram corresponding to the model. We fitted the model in the software R (R Core Team, 2014) and rstan (Stan Development Team, 2014), by running 3 chains for 7000 iterations, and discarding the first half of each as a 'warm up' phase, which allows the sampler to locate and sample from the probability distributions corresponding to the model. Stan code to fit the model is provided in 11.7.3. Convergence of the model was assessed using traceplots and the Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992). We conducted extensive model checking, by plotting the posterior predictive distribution against the observed data (Gelman, et al., 1996).

We conducted auxiliary analyses where we estimated the effects of gonadotropin dose on uterine receptivity separately in cleavage stage (day 2 or 3) and blastocyst (day 5) transfers. We reasoned that the effects of COS should be diminished in the latter, since the uterine environment has had more time to equilibrate.



	Outcomes	Covariates	
Submodel 1	Total Gonadotropins Dose *Log(dose)*	Age, Attempt No, Initial dose	
Submodel 2	Oocyte Yield *Count of eggs*	Downregulation protocol, Gonadotropin, Age, OPU practitioner, Diagnosis	Total Dose
Submodel 3	Fertilisation rate *Proportion of eggs*	Downregulation protocol, Gonadotropin, Age, Diagnosis, Partner Age, ICSI	Total Dose
Submodel 4	Embryo fragmentation *1 to 4 scale*	Downregulation protocol, Gonadotropin, Age, Partner Age, ICSI	Total Dose, Oocyte yield
Submodel 5	Number of embryos transferred *1 or 2*	Age, Attempt No	Oocyte yield, Fertilisation, Fragmentation
Submodel 6	Embryos implant *Yes/ No*	Downregulation protocol, Gonadotropin, Age, Diagnosis, Transfer practitioner	Total Dose, Oocyte yield, Fertilisation, Fragmentation, Number transferred
Submodel 7	Live birth event *Yes/ No*	Downregulation protocol, Gonadotropin, Age, Diagnosis, Transfer practitioner	Total Dose, Oocyte yield, Fertilisation, Fragmentation, Number transferred

Figure 40: Outcome variables and covariates in seven submodels in the model. Covariates in blue boxes are variables appearing as both response variables and covariates in the model.

11.4 Results

We identified 2968 eligible cycles on 2457 patients. Characteristics of the sample are shown in Table 18 and Table 19. The median (IQR) total dose of gonadotropins (IU) administered was 2250 (1581 to 3000).

11.4.1. Model checking

The convergence of the model parameters was deemed to be good on the basis of Gelman-Rubin statistics and traceplots (S Table 22 to S Table 30), and S Figure 8 to S Figure 15). When convergence is achieved, Gelman-Rubin statistics should be very close to 1 (Gelman and Rubin, 1992), and the traceplots should resemble overlaid dense scribbles. Plots of the model predictions against the observed responses indicated that the model closely resembled the multivariate distribution of the data (S Figure 4 to S Figure 7). On the basis of these checks, we were satisfied both that our algorithm successfully represented our intended model, and that our model provided a good fit to the data.

11.4.2. Estimates from the fitted model

The analysis produces a large number of estimates corresponding to the effects of different variables at different stages of treatment. While it is tempting to probe these for insights into the relationships between events in the cycle, we would stress that the present model was designed for a specific purpose: to validly identify the effects of COS upon downstream responses, with particular interest in effects on uterine receptivity. However, for completeness, we present regression tables corresponding to each response in the model in S Table 22 to S Table 30.

Variable		Summary
No of cycles started		2968
Number of egg collections (% per started)		2901 (98%)
Number of oocytes per cycle started		9 5 to 13 0 to 50
Number of cycles with fertilization attempted (% per started)		2867 (97%)
Method of fertilisation (per cycle with fertilisation attempt)		
	IVF	1203 (42%)
	ICSI	1664 (58%)
Number of cleaved embryos per cycle started		4 2 to 6 0 to 20
Total cleaved embryos		12 936
Day 2 embryo fragmentation grade (% per embryo)		
	1	340 (3%)
	2	2077 (16%)
	3	4027 (31%)
	4	6492 (50%)
Number of transfer procedures (% per started)		2505 (84%)
Number of embryos transferred per transfer		
	1	1049 (58%)
	2	1456 (42%)
Number of cycles where >=1 embryo implanted (% per cycle started)		1101 (37%)
Number of live birth events		811 (27%)
Attempt Number		
	1st	2136 (72%)
	2nd	661 (22%)
	3rd	147 (5%)
	4th	4 (0%)
	5th	20 (1%)

Table 18: Characteristics of the sample. Frequency (%) for categorical variables. Median, interquartile range and range for continuous variables. Continued in Table 19

Variable		Summary
Age (yrs)		33
		30 to 36
		21 to 43
Partner Age (yrs)		35
		32 to 39
		19 to 72
Initial Dose of gonadotropins (IU)		188
		150 to 300
		75 to 450
Total Dose of gonadotropins (IU)		2250
		1581 to 3000
		224 to 7650
Downregulation protocol		
	Antagonist	1612 (54%)
	Long Agonist	1356 (46%)
	OPU practitioner	
	1 (Practitioners with < 30 procedures)	108 (4%)
	2	214 (7%)
	3	380 (13%)
	4	104 (4%)
	5	241 (8%)
	6	133 (4%)
	7	258 (9%)
	8	99 (3%)
	9	569 (19%)
	10	661 (22%)
	11	201 (7%)
Transfer Practitioner 26 randomly allocated		
	1 (Practitioners with < 20 procedures)	46 (2%)
	2	141 (6%)
	3	166 (7%)
	4	136 (5%)
	5	164 (7%)
	6	185 (7%)
	7	125 (5%)
	8	191 (8%)
	9	36 (1%)
	10	120 (5%)
	11	393 (16%)
	12	580 (23%)
	13	198 (8%)
Diagnosis		
	Tubal	460 (15%)
	Ovulation Failure	37 (1%)
	Uterine Problem	29 (1%)
	Unexplained	809 (27%)
	Male factor	1506 (51%)
	Endometriosis	117 (4%)
	Anovulation	418 (14%)

Table 19: Continuation of Table 18.

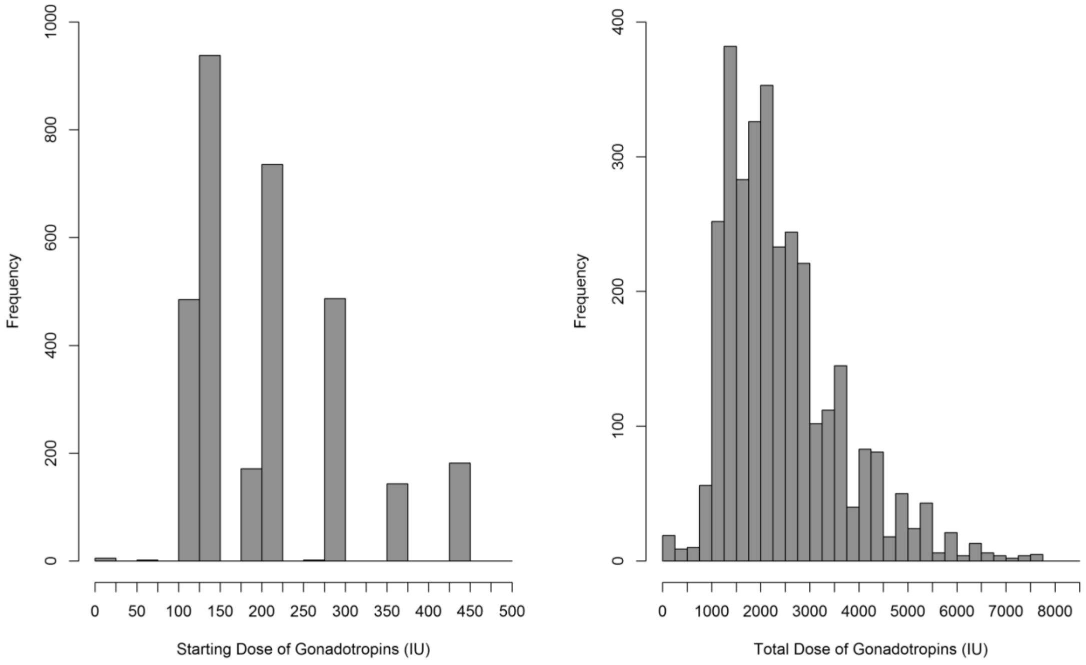


Figure 41: Distribution of starting dose and of total dose of gonadotropins (IU) in the dataset.

11.4.3. COS effects in the IVF cycle

Since the effects of COS are complex and nonlinear, they cannot be summarised by a single number. Instead, they are best understood by way of visual representations. The estimated direct effects of COS upon downstream response variables are displayed in Figure 42, with a shaded band representing +/-1 standard error. These estimates are adjusted for model covariates, and are also corrected for unmeasured confounders (such as levels of ovarian reserve). They represent dose effects after averaging over the other variables; the dosing thresholds described here therefore correspond to hypothetical 'average' patients. The analysis indicates that higher total doses increase oocyte yield up to a point (around 1300IU), after which the effect becomes deleterious. There is some suggestion that fertilisation rate dips as dose increases in the low range (up to around 1100IU), but then increases. This is the inverse of the relationship between dose and oocytes (however, the fertilisation curve may be misleading, and we explain why in the discussion). Our estimated curve corresponding to dose effects on embryo fragmentation suggests lower embryo quality with increasing dose, but is quite imprecise. We find that increasing gonadotropin dose negatively affects embryo implantation, with a downwards trend across the dose range. A decreasing probability of an implanted embryo being sustained to term is observed up to a total dose of around 1900 IU, but is relatively flat thereafter.

11.4.4. COS effects in cleavage and blastocyst stage transfers

Figure 43 shows the estimated dose effects on implantation and subsequent live birth event in cleavage stage (1580 cycles, 63%) and blastocyst (925 cycles, 37%) transfers. As expected, the dose effect on implantation appears to be more severe in cleavage compared to blastocyst transfers, albeit with considerable uncertainty. However, the patterns of effect on live birth are less distinct. This was evaluated in smaller numbers of cycles however (578 and 523, respectively), since it applies only to those cycles where embryos implant.

11.5 Discussion

We have conducted a multistage modelling study to determine the effects of COS throughout the IVF cycle. Ours is the first study on this topic to use statistical methodology capable of distinguishing between the various mechanisms by which COS affects treatment outcome. Our findings support and augment the limited evidence available to date, which has directly and indirectly suggested that COS adversely affects endometrial angiogenesis (Haouzi, et al., 2009, Healy, et al., 2010, Kalra, et al., 2011). Our analysis suggests that COS reduces the likelihood that transferred embryos will implant, with evidence of a nonlinear overall dose-response relationship. The analysis also indicates that COS reduces the chance of an implanted embryo being carried to term. The probability appears to decrease as total dose increases, up to a dose of around 1900IU. Beyond this, we do not see evidence of any further reduction with increasing dose.

Our model suggests increased oocyte yields with higher doses up to around 1300IU, and

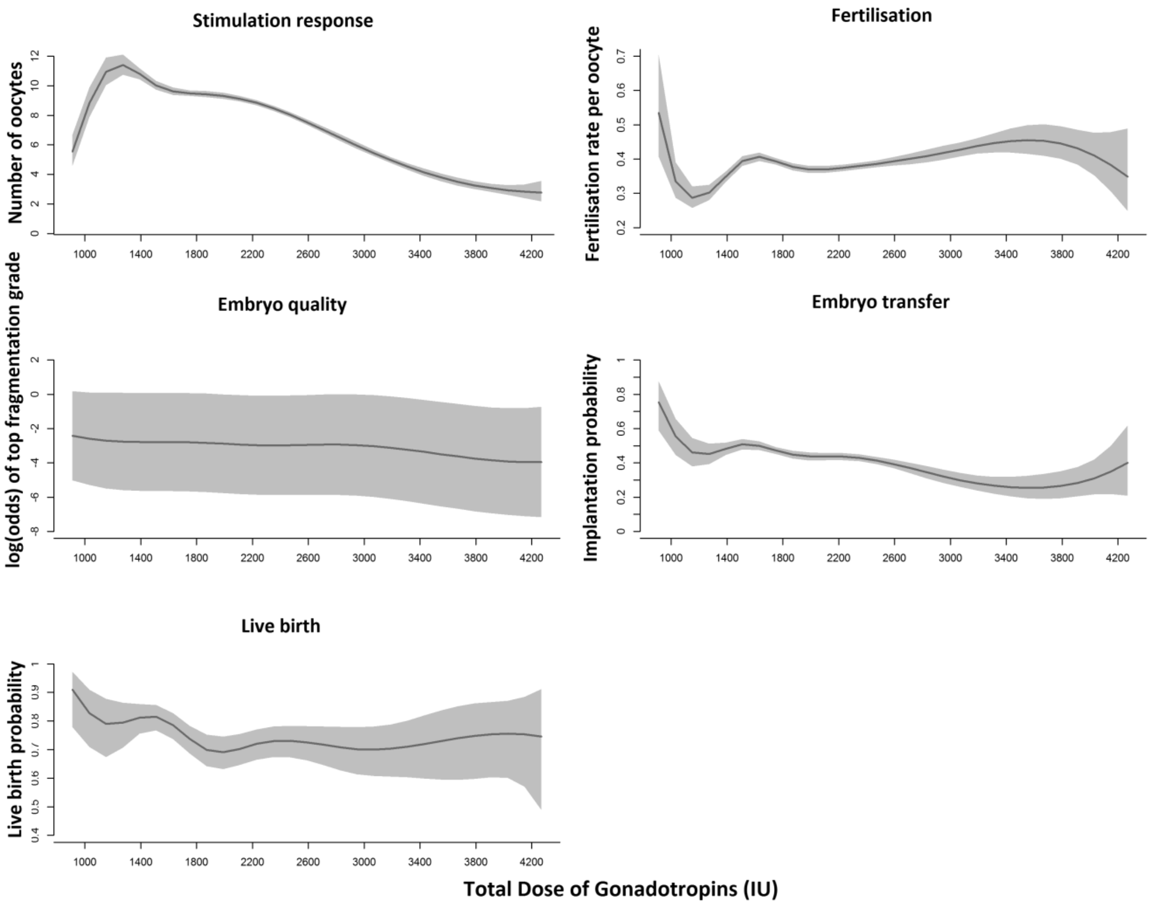


Figure 42: Estimated direct effects (+/-1SE) of total gonadotropin dose on responses in the model. Estimates are adjusted for confounding arising from measured and unmeasured covariables.

reduced yields beyond this. As described in the Statistical Analysis section, this estimate is adjusted for unmeasured confounders, such as ovarian reserve. This finding appears to be consistent with previous studies which have suggested a dose-response relationship between initial gonadotropin dose and stimulation response at low doses (Arce, et al., 2014). Initial dose effects have previously been shown to be to be modest (Sterrenburg, et al., 2011). The present analysis suggests that dose effects become more defined over the stimulation period. While we also found an inverse relationship between dose and fertilisation (reduction in the fertilisation rate up to a dose of around 1100IU, followed by an increase), we are less confident that this can be attributed to COS *per se*. This is due to the (technical) point that the number of oocytes appears as an 'offset' rather than as a covariate in the fertilisation submodel. A consequence of this is that direct dose effects on fertilisation are not clearly distinguished from knock on effects following from the effects of dose on oocyte yield. Accordingly, the apparent effect of dose on oocytes could be attributable to larger yields including more immature oocytes (for example). As such, the dose curve we present for fertilisation should be interpreted with caution. Since a similar pattern is evident in several of the submodels (a steeper change in response at low doses, which then plateaus or inverts), characteristics of the oocytes and resulting embryos, rather than physiological effects of COS, might be offered as the probable explanation. However, our model is able to distinguish dose effects from the effects ensuing from having an increased pool of oocytes, for all of response variables other than fertilisation rate.

There is some suggestion that gonadotropin dosing has physiological effects up to a threshold, which manifest both in stimulation response and also in uterine receptivity. This is consistent with the hypothesis, discussed by Maheshwari and Bhattacharya (2013), that negative uterine effects of COS are due to oestrogen produced by growing follicles. A biological limit on the number of follicles activated by increasing gonadotropin dose (as suggested, for example, by Sterrenburg and colleagues, (2011)) would then explain both the observed relationship with the number of oocytes obtained and the relationship with uterine receptivity. Our subgroup analyses according to day of transfer appear to support the hypothesis; the negative effects of COS on implantation may be diminished in

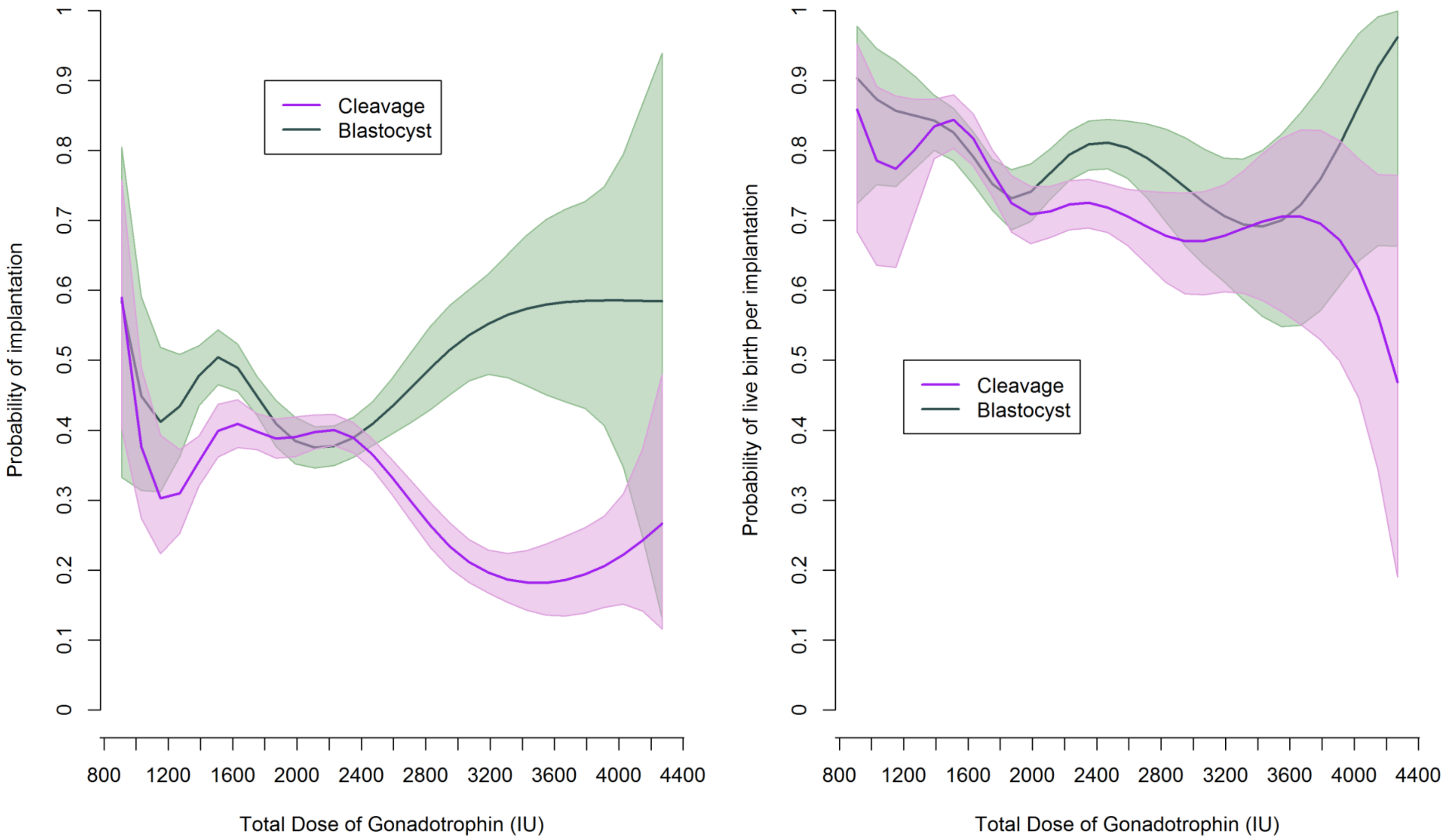


Figure 43: Estimated direct effects of gonadotropin on embryo implantation (left) and live birth event (right), adjusted for measured and unmeasured confounding. Shaded area = 1SE.

blastocyst compared to cleavage stage transfers. These latter analyses suffer from imprecision however.

Our findings broadly agree with previous studies; they support the case for treatment strategies to mitigate the disturbance caused by COS, such as elective cryopreservation of embryos. However, we would generally caution against using standard regression methods to answer mechanistic questions relating to IVF. IVF is a complex, multistage intervention, and special statistical methods are required to delineate the underlying causal network (Blalock, 1961, Terza, 1998). We would encourage quantitative researchers in this field to acquaint themselves with joint and structural modelling techniques, which can be used to estimate effects of interventions by different pathways in the presence of unmeasured confounding. That said, it is important to consider exactly how our analysis could be wrong, and what impact this would have on our conclusions. Any parametric analysis, whether it is a t-test or a complex multivariate regression like the one presented here, rests on the adequacy of the assumed model. We have confirmed that the model is consistent with the observed data, by simulating data from the fitted model and comparing to the actual responses. We have also conducted sensitivity analyses, where we relaxed the regularisation constraints used to fit the model, specifically by allowing greater variation in the model intercepts. Our dose estimates were robust to these investigations. Confounding always represents a major concern in non-randomised studies. Our approach is to adjust for confounding as far as possible by including covariates in the model, and to minimise unmeasured confounding by including correlated latent variables in the model (Skrondal and Rabe-Hesketh, 2004, Terza, 1998). This approach may not eliminate the problem entirely however. As such, we would recommend further studies of differing designs, to confirm that the effects we describe are independent of bias (ie: to adopt a triangulation approach, Lawlor, et al., 2016). For example, an instrumental variable analysis of RCT data (Emsley, et al., 2010), would provide a useful adjunct to the present study.

While the subjective embryo fragmentation grading we used as a measure of embryo quality in the model has been shown to be predictive of pregnancy (Stylianou, et al., 2012), it will nevertheless be subject to some measurement error. In fact, our model estimates suggest that embryos with higher fragmentation gradings are less likely to implant and to be sustained to term (S Table 27 and S Table 30). A disadvantage of transferring two embryos at once on implantation is also suggested (S Table 28). However, we conducted a sensitivity analysis where we additionally adjusted for the day of transfer (2, 3, or 5) in the implantation and live birth submodels. Clear advantages of higher grade embryos and of double embryo transfer on implantation were evident once we accounted for the fact that blastocysts are more likely to be transferred in singleton. A negative estimate for the effect of higher fragmentation grade on live birth was still observed however. Accordingly, we would repeat our advice against interpreting the other model parameters (Westreich and Greenland, 2013).

We have not been able to investigate whether the dose effects we have described here vary according to the patient’s ovarian reserve in this study. While ovarian reserve tests are used for initial dose and protocol selection in our centre, the measurements themselves are not recorded in the database. If deleterious implantation effects are due to follicular development, then it is plausible that patients with low or high reserve might be differentially affected. This remains to be established; the present analysis describes only average effects of dose in the cohort. It should be noted that the question of whether or not dose effects differ according to ovarian reserve is not the same as the question of whether or not apparent dose effects are really attributable to confounding due to ovarian reserve. We are agnostic about the former, but our method rules out the latter.

Numerous RCTs of elective frozen transfers are underway, and until these studies are completed, the clinical effectiveness and safety of this strategy is unknown (Wong, et al., 2017). Our findings suggest that, if these trials are negative, the cause will lie in effects of cryopreservation on embryos. The implications for patient safety must also be considered. In the interim, clinicians should be aware that the likelihood of successful implantation and gestation appears to be highest with low cumulative exposure to gonadotropins. Wherever possible, a strategy of transferring blastocyst stage embryos appears to offer some protection against negative effects of COS.

11.6 References for Chapter 11.

Arce JC, Andersen AN, Fernandez-Sanchez M, Visnova H, Bosch E, Garcia-Velasco JA, Barri P, De Sutter P, Klein BM, Fauser BCJM. Ovarian response to recombinant human follicle-stimulating hormone: a randomized, antimullerian hormone-stratified, dose-response trial in women undergoing in vitro fertilization/intracytoplasmic sperm injection. *Fertil Steril* 2014;102: 1633-U1456.

Blalock HM. Evaluating the Relative Importance of Variables. *Am Sociol Rev* 1961;26: 866-874.

Dunson DB. Bayesian latent variable models for clustered mixed outcomes. *J R Stat Soc Series B Stat Methodol* 2000;62: 355-366.

Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* 2003;59: 521-530.

Emsley R, Dunn G, White IR. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Stat Methods Med Res* 2010;19: 237-270.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis, 2014. CRC press Boca Raton, FL.

Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat Sinica* 1996;6: 733-760.

Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci* 1992: 457-472.

Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. *Stat Model* 2009;9: 173-197.

Haouzi D, Assou S, Mahmoud K, Tondeur S, Reme T, Hedon B, De Vos J, Hamamah S. Gene expression profile of human endometrial receptivity: comparison between natural and stimulated cycles for the same patients. *Hum Reprod* 2009;24: 1436-1445.

Harbottle S, Hughes C, Cutting R, Roberts S, Brison D, Embryologists AC, BFS ABFS. Elective Single Embryo Transfer: an update to UK Best Practice Guidelines. *Hum Fertil* 2015;18: 165-183.

Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis, 2015. Springer.

Healy DL, Breheny S, Halliday J, Jaques A, Rushford D, Garrett C, Talbot JM, Baker HWG. Prevalence and risk factors for obstetric haemorrhage in 6730 singleton births after assisted reproductive technology in Victoria Australia. *Hum Reprod* 2010;25: 265-274.

Heckman JJ. Common Structure of Statistical-Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Ann Econ Soc Meas* 1976;5: 475-492.

Heckman JJ. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* 1978;46: 931-959.

Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *Am J Epidemiol*. 2002;155: 176-184.

Kalra SK, Ratcliffe SJ, Milman L, Gracia CR, Coutifaris C, Barnhart KT. Perinatal morbidity after in vitro fertilization is lower with frozen embryo transfer. *Fertil Steril* 2011;95: 548-553.

Maheshwari A, Bhattacharya S. Elective frozen replacement cycles for all: ready for prime time? *Hum Reprod* 2013;28: 6-9.

Maheshwari A, Pandey S, Shetty A, Hamilton M, Bhattacharya S. Obstetric and perinatal outcomes in singleton pregnancies resulting from the transfer of frozen thawed versus fresh embryos generated through in vitro fertilization treatment: a systematic review and meta-analysis. *Fertil Steril* 2012;98: 368-+.

R Core Team. R: A language and environment for statistical computing. 2014. R Foundation for Statistical Computing, Vienna, Austria.

Raziel A, Schachter M, Friedler S, Ron-El R. Outcome of IVF pregnancies following severe OHSS. *Reprod Biomed Online* 2009;19: 61-65.

Rubin DB. Inference and Missing Data. *Biometrika* 1976;63: 581-590.

Shapiro B, Daneshmand S, Bedient C, Garner F. Comparison of birthweights in patients randomly assigned to fresh or frozen-thawed embryo transfer. Hum Reprod Conference: 32nd Annual Meeting of the European Society of Hum Reprod and EmbryologyFinland 2016;31: i367.

Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models, 2004. Crc Press.

Stan Development Team. RStan: the R interface to Stan, Version 2.5.0. 2014.

Steele F, Washbrook E. Discrete-time Event History Analysis. In Centre for Multilevel Modelling UoB (ed). 2013.

Sterrenburg MD, Veltman-Verhulst SM, Eijkemans MJC, Hughes EG, Macklon NS, Broekmans FJ, Fauser BCJM. Clinical outcomes in relation to the daily dose of recombinant follicle-stimulating hormone for ovarian stimulation in in vitro fertilization in presumed normal responders younger than 39 years: a meta-analysis. *Hum Reprod Update* 2011;17: 184-196.

Steward RG, Lan L, Shah AA, Yeh JS, Price TM, Goldfarb JM, Muasher SJ. Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: an analysis of 256,381 in vitro fertilization cycles. *Fertil Steril* 2014;101: 967-973.

Stylianou C, Critchlow D, Brison DR, Roberts SA. Embryo morphology as a predictor of IVF success: An evaluation of the proposed UK ACE grading scheme for cleavage stage embryos. *Hum Fertil* 2012;15: 11-17.

Terza JV. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *J Econom* 1998;84: 129-154.

Walker AM. Confounding by indication. *Epidemiology*. 1996;7(4):335-6.

Weinerman R, Mainigi M. Why we should transfer frozen instead of fresh embryos: the translational rationale. *Fertil Steril* 2014;102: 10-18.

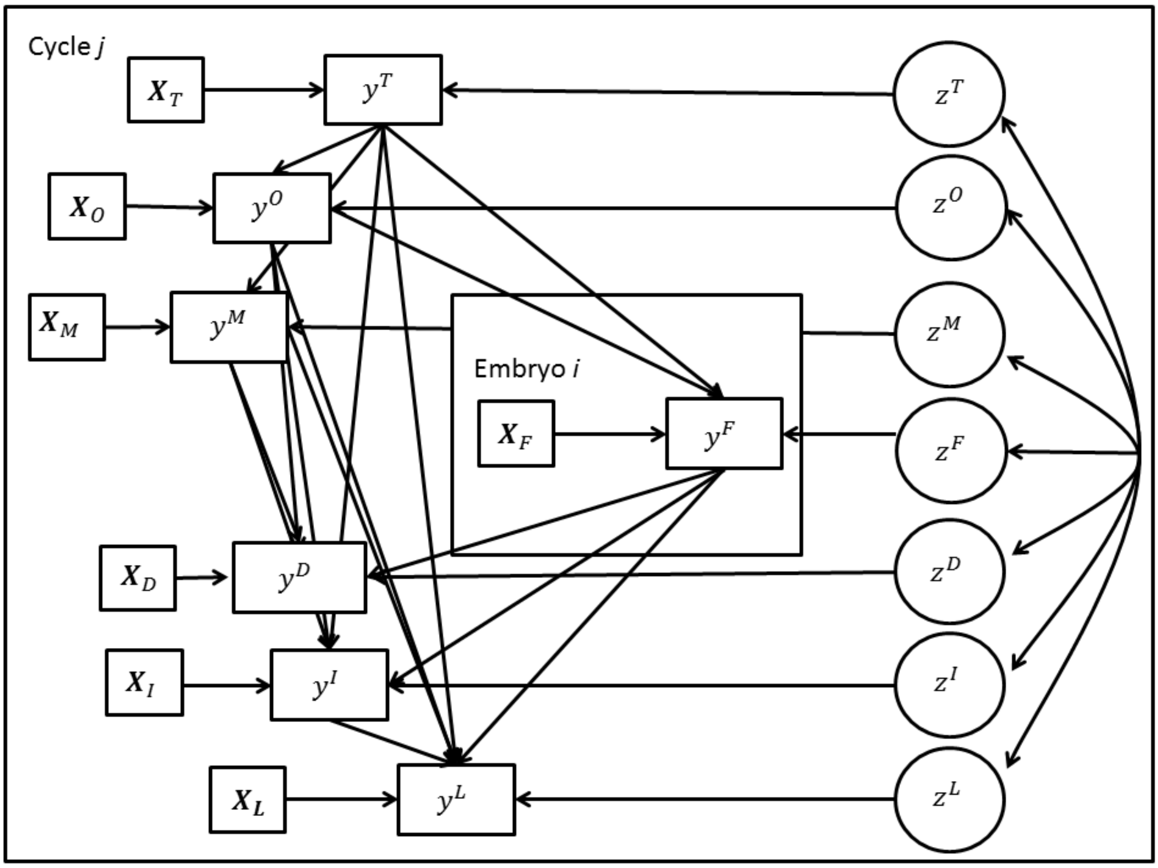
Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol* 2013;177: 292-298.

Wong KM, van Wely M, Mol F, Repping S, Mastenbroek S. Fresh versus frozen embryo transfers in assisted reproduction. *Cochrane Db Syst Rev* 2017.

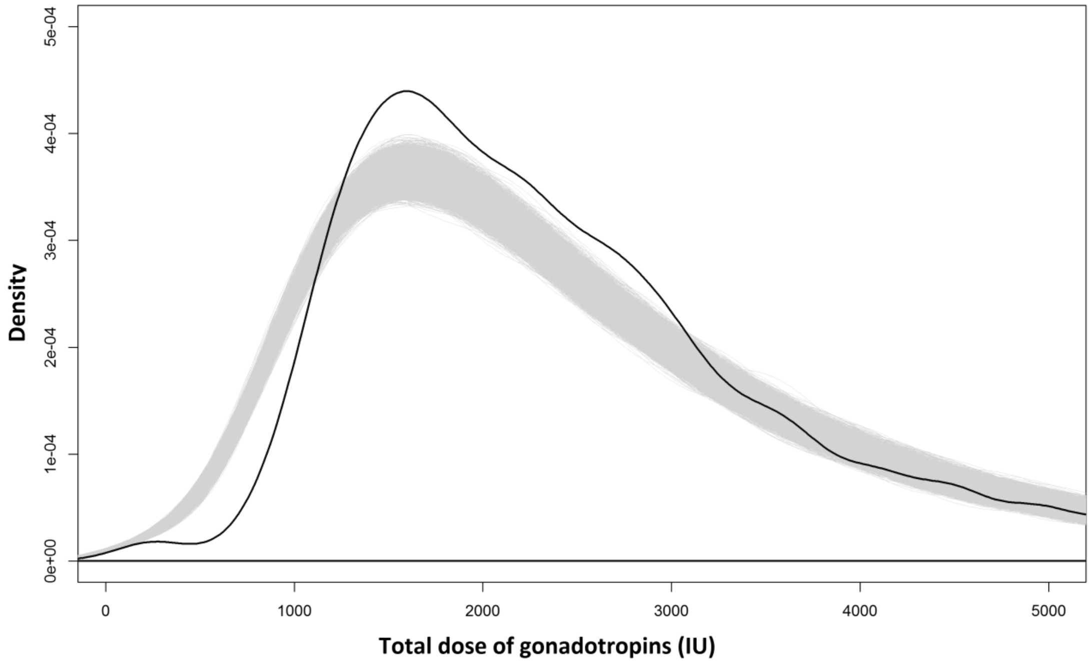
11.7 Supplementary Material for Chapter 11.

11.7.1. Supplementary Figures for Chapter 11.

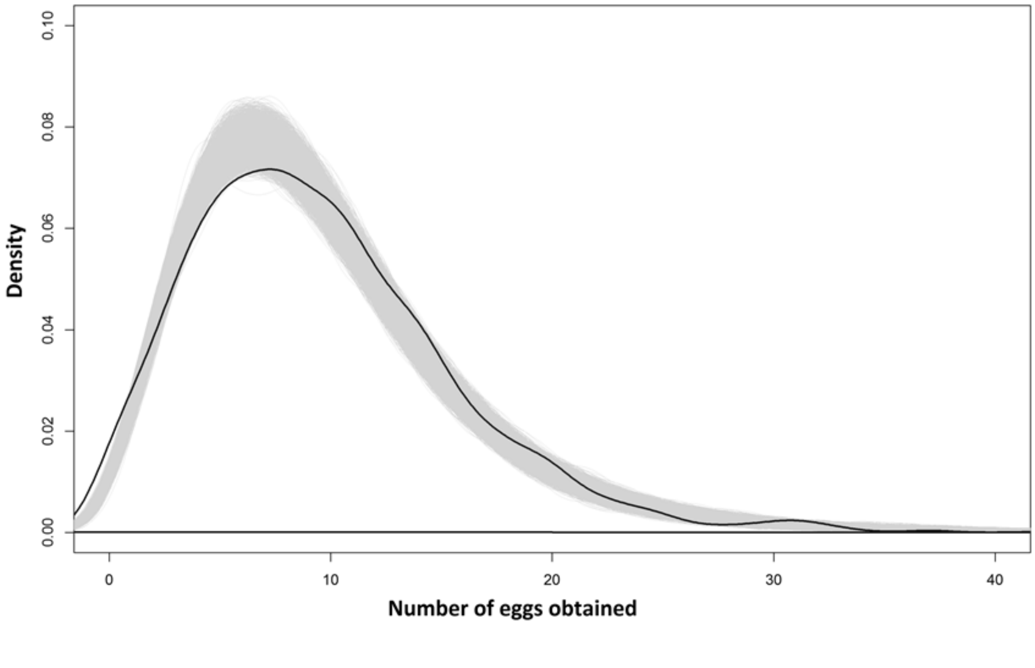
[Beginning on next page]



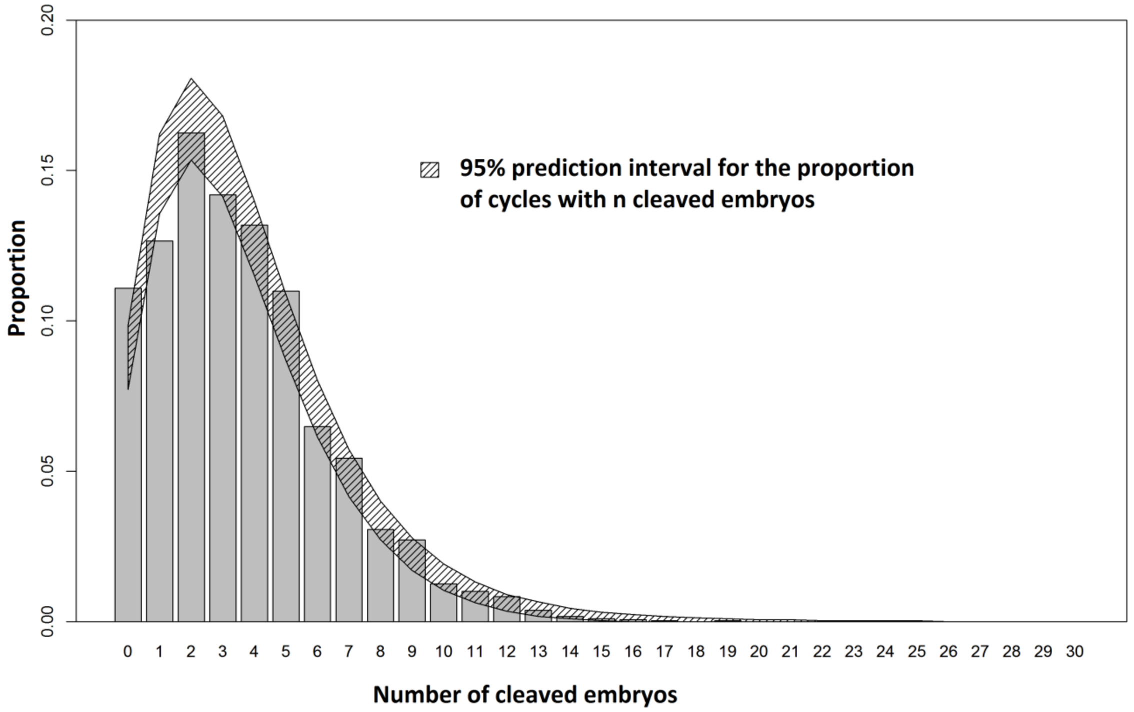
S Figure 3: Path diagram showing the causal structure implied by the model. Arrows point from causal antecedents to consequents. Mathematical notation is presented in S File 1.



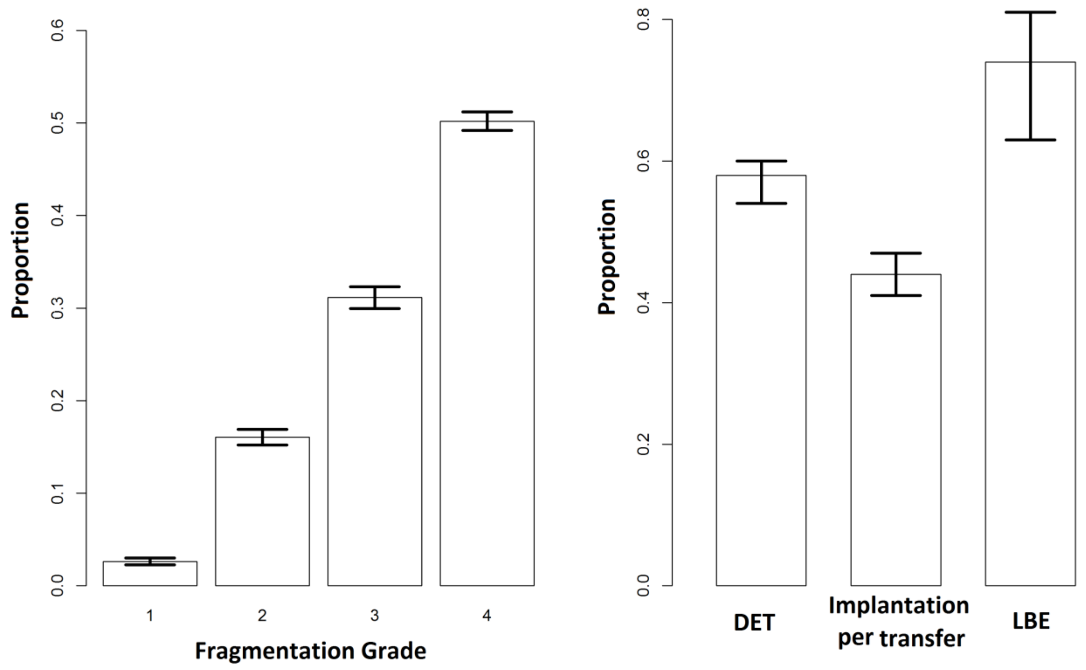
S Figure 4: Model checking. Observed distribution of total doses of gonadotropin (dark line) with simulated distributions from the posterior predictive distribution (clumped light grey lines).



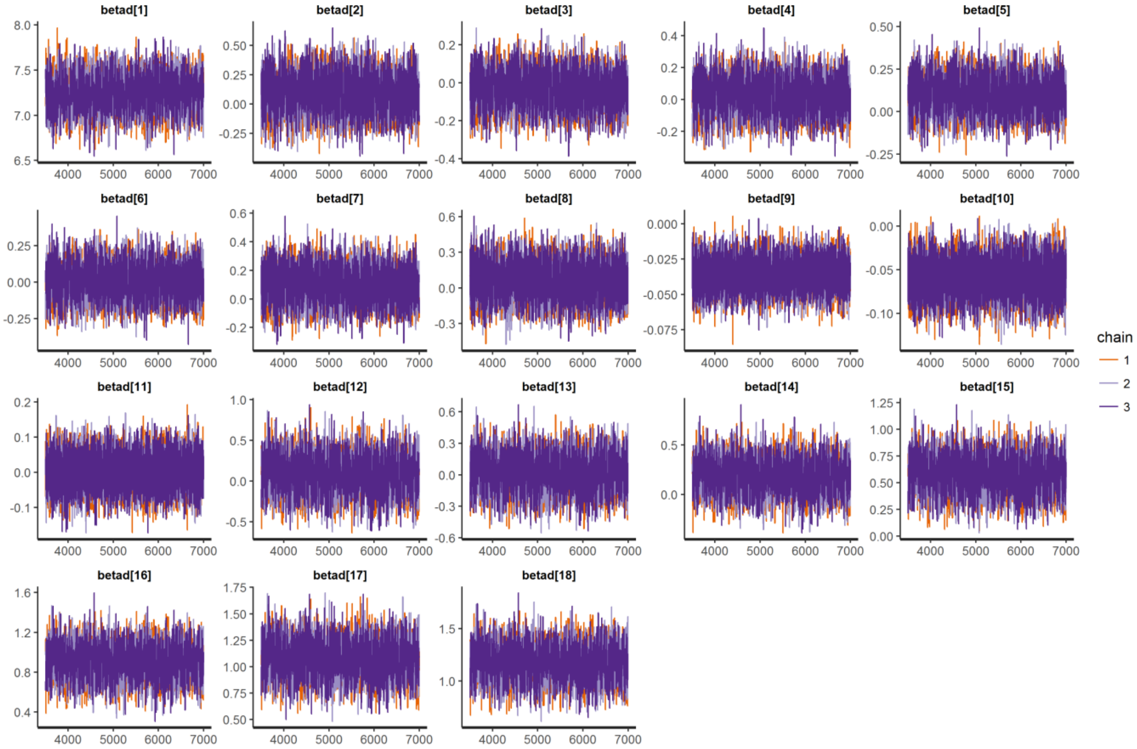
S Figure 5: Model checking. Observed distribution of number of eggs obtained (dark line) with simulated distributions from the posterior predictive distribution (clumped light grey lines).



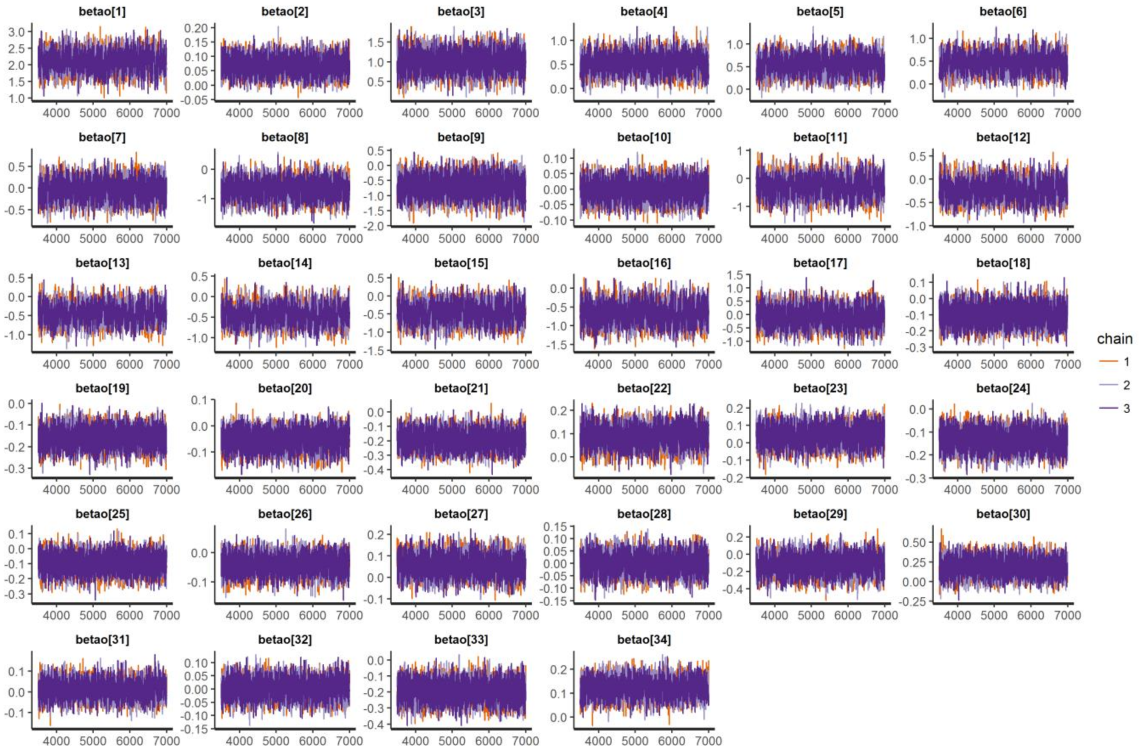
S Figure 6: Model checking. Observed distribution of number of cleaved embryos (successful fertilisations)



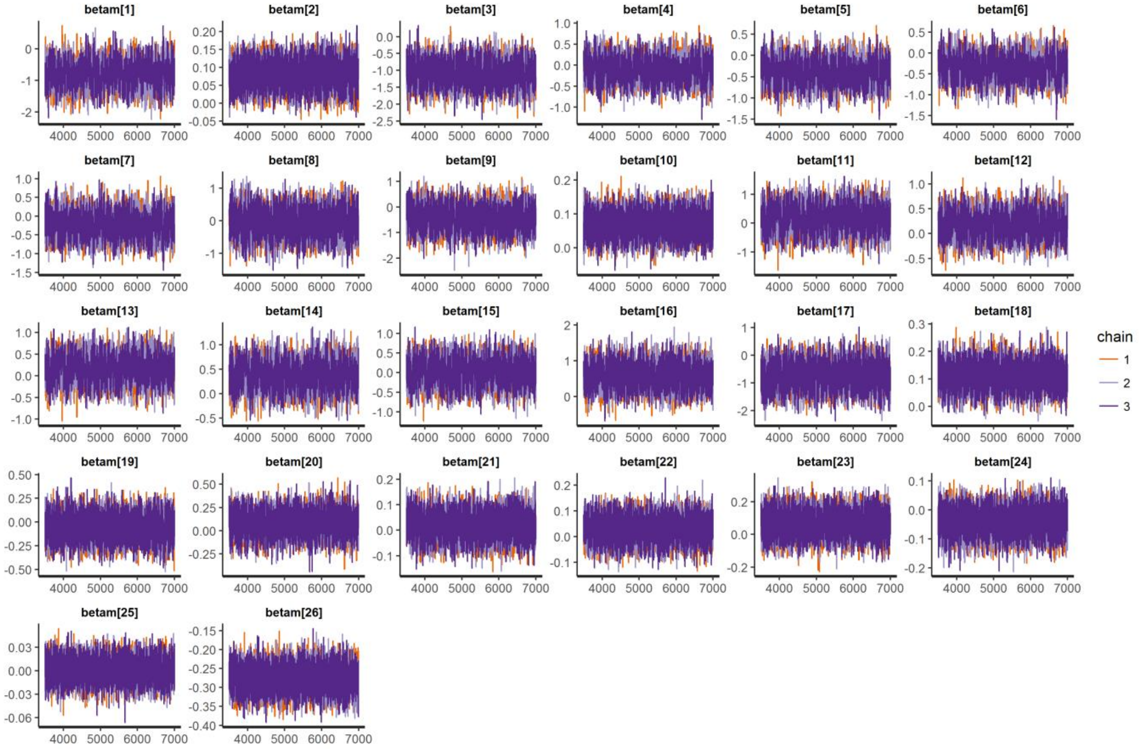
S Figure 7: Model checking for fragmentation grade, double embryo transfer (DET), implantation and live birth event (LBE). Bar heights show observed proportions in the dataset. 'Error bars' correspond to 95% prediction intervals for the proportions drawn from the posterior predictive distribution.



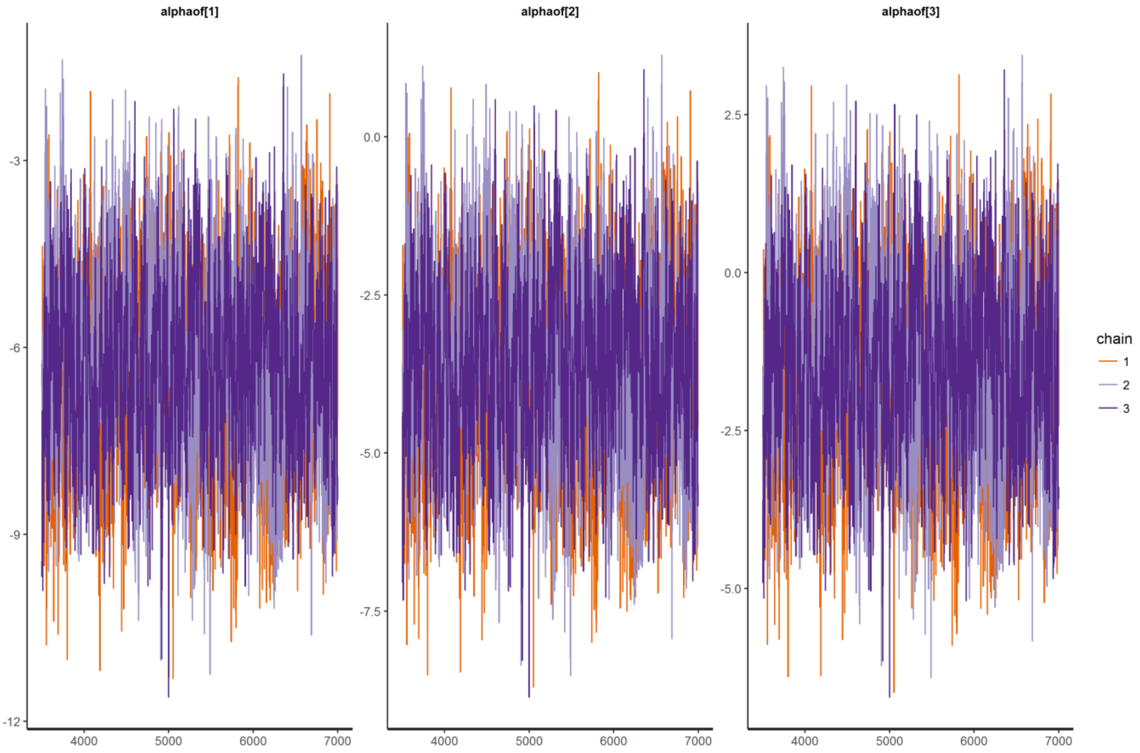
S Figure 8: Traceplots for dose submodel



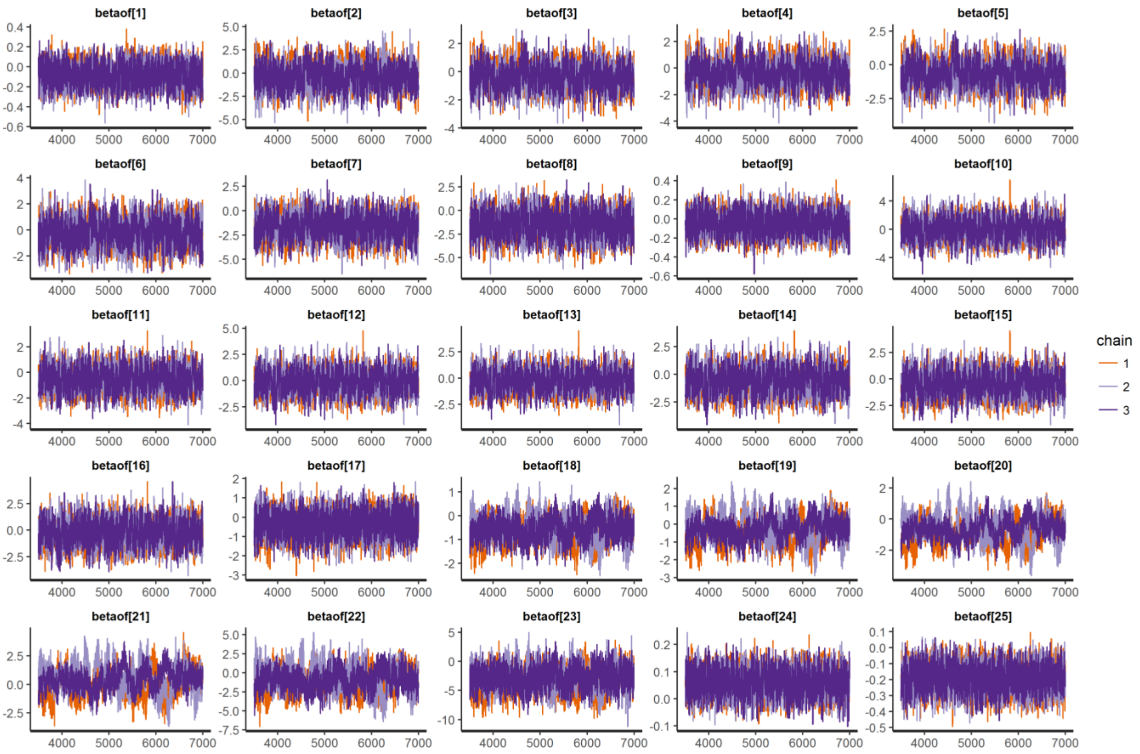
S Figure 9: Traceplots for number of oocytes submodel



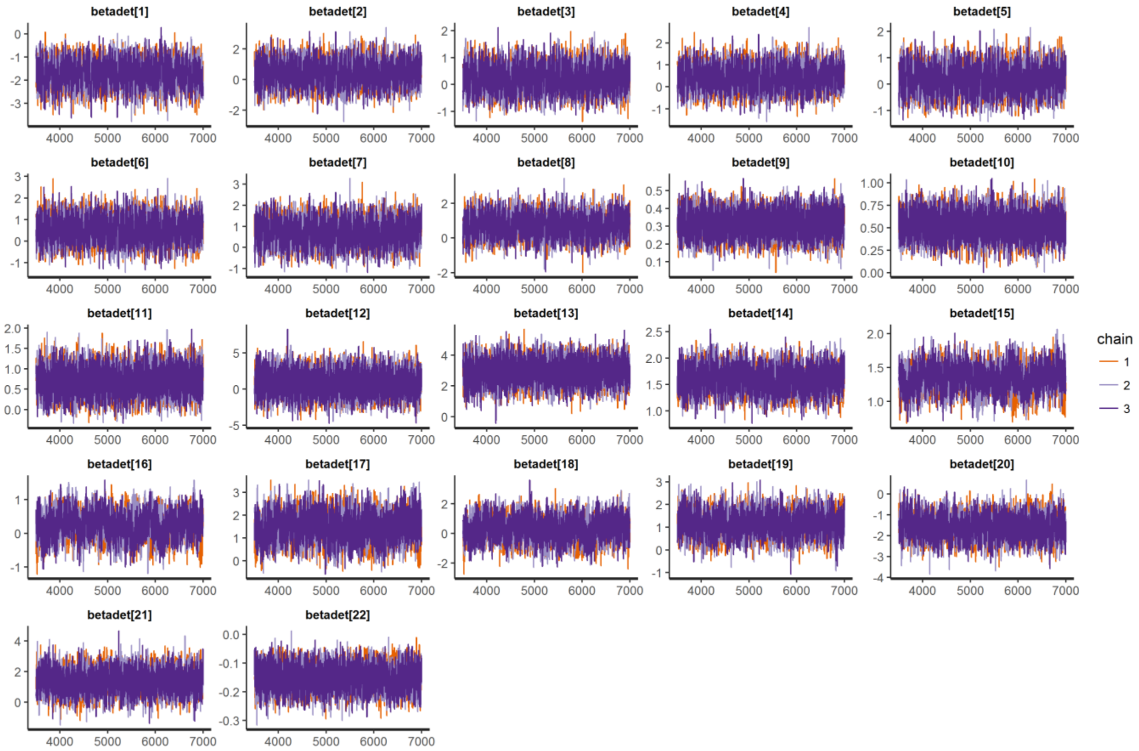
S Figure 10: Traceplots for fertilisation rate submodel



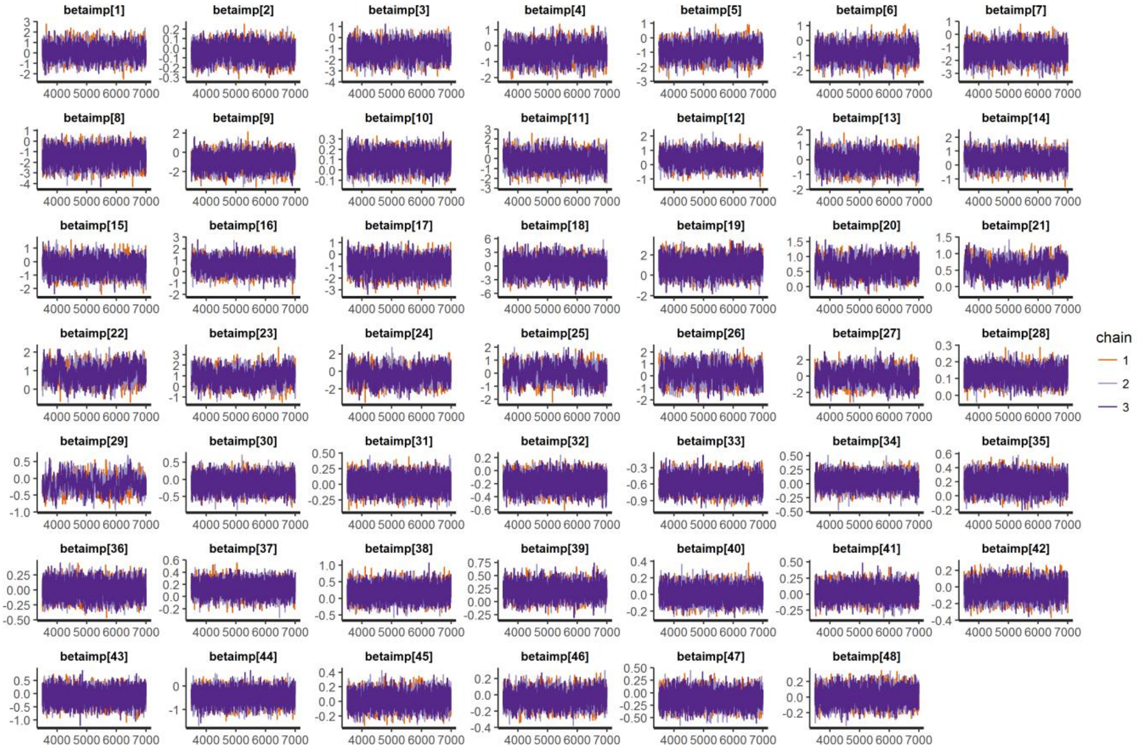
S Figure 11: Traceplots for threshold parameters in embryo fragmentation submodel



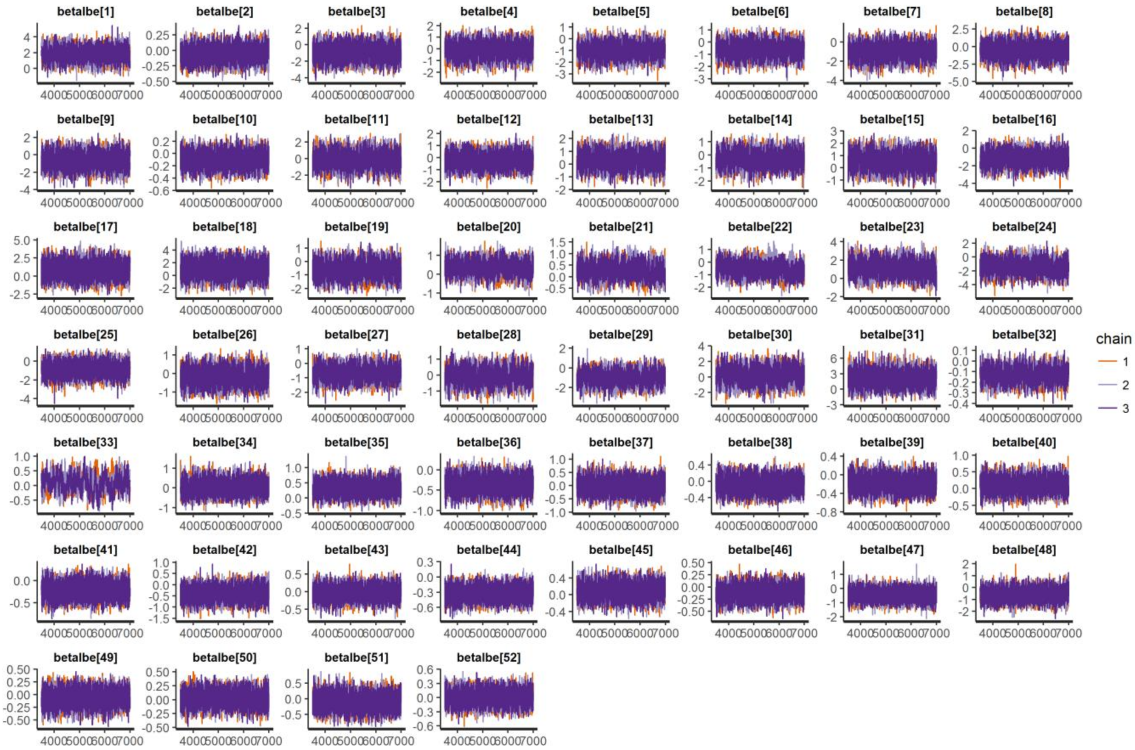
S Figure 12: Traceplots for regression coefficients in embryo fragmentation submodel.



S Figure 13: Traceplots for double embryo transfer submodel.



S Figure 14: Traceplots for embryo implantation submodel



S Figure 15: Traceplots for live birth event submodel

11.7.2. Supplementary Tables for Chapter 11.

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Log(Total Dose) submodel	Linear regression of log (dose)				
Intercept	Mean log dose averaging over other covariates	7.254	6.882 to 7.618	2176	1.00
Age Spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.094	-0.211 to 0.389	2854	1.00
2		-0.037	-0.215 to 0.138	2974	1.00
3		0.026	-0.187 to 0.230	2705	1.00
4		0.101	-0.098 to 0.294	2727	1.00
5		0.005	-0.216 to 0.223	2988	1.00
6		0.097	-0.130 to 0.323	3349	1.00
7		0.054	-0.230 to 0.342	4312	1.00
Attempt No: 1st	Categorical variable w/ 4 levels				
2nd		-0.036	-0.057 to -0.015	10500	1.00
3rd		-0.059	-0.100 to -0.019	10500	1.00
4th or 5th	Combined due to low numbers	0.001	-0.093 to 0.097	10500	1.00
Initial dose spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.084	-0.333 to 0.515	2639	1.00
2		0.017	-0.306 to 0.344	2722	1.00
3		0.184	-0.131 to 0.507	2643	1.00
4		0.577	0.260 to 0.901	2609	1.00
5		0.905	0.590 to 1.228	2631	1.00
6		1.071	0.752 to 1.405	2709	1.00
7		1.170	0.862 to 1.488	2574	1.00

S Table 22: Characteristics of the log(total dose of gonadotropin) submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to changes in log(total dose) as each covariate varies.

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Number of oocytes submodel	Poisson regression				
Intercept	Mean log oocytes per cycle	2.124	1.534 to 2.676	1616	1.00
Long Agonist Protocol (vs Antagonist)	Binary indicator variable	0.069	0.012 to 0.128	3662	1.00
Total dose spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.998	0.525 to 1.492	2854	1.00
2		0.532	0.161 to 0.910	1862	1.00
3		0.541	0.169 to 0.933	1624	1.00
4		0.469	0.087 to 0.865	1559	1.00
5		-0.044	-0.477 to 0.408	1953	1.00
6		-0.663	-1.250 to -0.091	2311	1.00
7		-0.694	-1.318 to -0.062	2702	1.00
HMG (vs FSH)	Binary indicator variable	-0.003	-0.063 to 0.056	3712	1.00
Age spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.294	-0.969 to 0.403	1577	1.00
2		-0.231	-0.624 to 0.184	1756	1.00
3		-0.446	-0.920 to 0.039	1517	1.00
4		-0.449	-0.887 to 0.002	1578	1.00
5		-0.466	-0.979 to 0.057	1682	1.00
6		-0.681	-1.195 to -0.158	2049	1.00
7		-0.062	-0.739 to 0.603	2229	1.00
OPU practitioner (vs 10)	Categorical variable w 11 levels. 10 used as reference category due to highest frequency.				
1 (Fewer than 30 ops)		-0.099	-0.215 to 0.015	5499	1.00
2		-0.158	-0.242 to -0.073	4818	1.00
3		-0.053	-0.124 to 0.015	3292	1.00
4		-0.189	-0.311 to -0.075	4384	1.00
5		0.087	0.005 to 0.169	1415	1.01
6		0.040	-0.068 to 0.147	1005	1.00
7		-0.135	-0.215 to -0.054	1937	1.00
8		-0.094	-0.210 to 0.019	5535	1.00
9		-0.040	-0.101 to 0.020	3616	1.00
11		0.056	-0.032 to 0.147	715	1.01
Diagnosis	Seven binary indicators				
Tubal disease		0.000	-0.074 to 0.074	3117	1.00
Ovulation Failure		-0.119	-0.327 to 0.090	1206	1.01
Uterine problem		0.191	-0.020 to 0.401	3083	1.00
Unexplained		0.013	-0.067 to 0.095	3080	1.00
Male factor		0.004	-0.064 to 0.071	3153	1.00
Endometriosis		-0.192	-0.310 to -0.077	4976	1.00
Anovulation		0.122	0.050 to 0.196	2077	1.00

S Table 23: Characteristics of the number of oocytes submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to log(yield ratios). Yield ratios indicate the relative change in the number of oocytes as each covariate varies.

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Fertilisation submodel	Poisson with log(oocytes) as offset variable				
Intercept	Mean log(cleaved embryos) averaging over other covariates	-0.795	-1.566 to -0.011	1709	1.00
Long Agonist Protocol (vs Antagonist)	Binary indicator variable	0.083	0.010 to 0.152	3966	1.00
Total dose spline	Spline with knots at quintiles. Not directly interpretable.				
1		-1.084	-1.780 to -0.386	2610	1.00
2		-0.097	-0.638 to 0.445	1558	1.00
3		-0.398	-0.956 to 0.168	1404	1.00
4		-0.347	-0.899 to 0.210	1296	1.00
5		-0.223	-0.855 to 0.413	1678	1.00
6		-0.047	-0.836 to 0.721	2036	1.00
7		-0.420	-1.331 to 0.440	2455	1.00
HMG (vs FSH)	Binary indicator variable	0.070	-0.002 to 0.140	4499	1.00
Age spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.155	-0.688 to 1.015	2051	1.00
2		0.176	-0.308 to 0.669	2241	1.00
3		0.149	-0.432 to 0.755	2006	1.00
4		0.370	-0.165 to 0.918	2015	1.00
5		0.007	-0.614 to 0.649	2203	1.00
6		0.581	-0.072 to 1.245	2828	1.00
7		-0.729	-1.620 to 0.160	2999	1.00
Diagnosis	Seven binary indicators				
Tubal disease		0.110	0.021 to 0.200	4750	1.00
Ovulation Failure		-0.048	-0.309 to 0.200	10500	1.00
Uterine problem		0.081	-0.168 to 0.320	5751	1.00
Unexplained		0.020	-0.079 to 0.116	4398	1.00
Male factor		0.030	-0.063 to 0.123	4712	1.00
Endometriosis		0.064	-0.080 to 0.209	10500	1.00
Anovulation		-0.041	-0.129 to 0.047	4210	1.00
Partner Age (SDs)	Standardised, linear term	0.001	-0.028 to 0.029	10500	1.00
ICSI (vs IVF)	Binary indicator variable	-0.275	-0.341 to -0.209	10500	1.00

S Table 24: Characteristics of the fertilisation submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to log(rate ratios). Rate ratios indicate the relative change in the rate of cleaved embryos per oocyte as each covariate varies.

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Embryo Fragmentation submodel	Cumulative logit model				
1st threshold parameter	Odds of grade 1	-6.309	-9.077 to -3.481	380	1.01
2nd threshold parameter	Odds of grade 2 or less	-3.643	-6.414 to -0.837	380	1.01
3rd threshold parameter	Odds of grade 3 or less	-1.524	-4.294 to 1.294	379	1.01
Long Agonist Protocol (vs Antagonist)	Binary indicator variable	-0.083	-0.300 to 0.136	1816	1.00
Total dose spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.337	-2.826 to 2.191	631	1.01
2		-0.385	-2.060 to 1.364	720	1.01
3		-0.354	-2.139 to 1.428	729	1.00
4		-0.669	-2.425 to 1.114	714	1.00
5		-0.204	-2.163 to 1.744	1365	1.00
6		-1.652	-4.025 to 0.677	957	1.00
7		-1.511	-4.118 to 1.025	1556	1.00
HMG (vs FSH)	Binary indicator variable	-0.048	-0.269 to 0.174	2532	1.00
Age spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.007	-2.916 to 2.903	1689	1.00
2		-0.505	-2.326 to 1.278	1849	1.00
3		-0.108	-2.225 to 2.009	1298	1.00
4		-0.243	-2.228 to 1.732	1184	1.00
5		-0.568	-2.702 to 1.605	1152	1.00
6		-0.436	-2.580 to 1.721	1019	1.00
7		-0.241	-2.479 to 2.026	1497	1.00
Oocyte spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.363	-1.592 to 0.839	1261	1.00
2		-0.586	-1.606 to 0.423	124	1.04
3		-0.315	-1.680 to 1.037	99	1.05
4		-0.678	-2.301 to 0.875	80	1.06
5		0.362	-1.907 to 2.653	86	1.06
6		-0.936	-4.014 to 2.132	117	1.04
7		-2.875	-6.918 to 0.997	211	1.02
Partner Age (SDs)	Standardised, linear term	0.059	-0.033 to 0.151	3388	1.00
ICSI (vs IVF)	Binary indicator variable	-0.193	-0.354 to -0.036	3201	1.00

S Table 25: Characteristics of the embryo fragmentation submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to log(odds ratios). Odds ratios indicate the relative change in the odds of the embryo having a higher, rather than a lower grade as each covariate varies.

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Double embryo transfer submodel	Probit regression model				
Intercept	Mean Z score, after averaging over other variables	-1.711	-2.787 to -0.683	2136	1.00
Age spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.423	-1.039 to 1.949	2619	1.00
2		0.230	-0.648 to 1.158	2674	1.00
3		0.430	-0.587 to 1.494	2371	1.00
4		0.169	-0.778 to 1.167	2392	1.00
5		0.544	-0.527 to 1.663	2656	1.00
6		0.718	-0.378 to 1.838	2772	1.00
7		0.718	-0.475 to 1.936	2710	1.00
Attempt number: 1st	Categorical variable w/ 4 levels				
2nd		0.314	0.182 to 0.444	10500	1.00
3rd		0.525	0.255 to 0.804	3094	1.00
4th or 5th	Combined due to low numbers	0.684	0.071 to 1.336	4665	1.00
Number of cleaved embryos spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.812	-2.251 to 3.979	7749	1.00
2		2.931	1.477 to 4.376	7094	1.00
3		1.603	1.147 to 2.040	1579	1.00
4		1.341	0.966 to 1.704	396	1.01
5		0.221	-0.532 to 0.980	362	1.01
6		1.472	0.293 to 2.639	793	1.00
7		0.313	-1.153 to 1.797	656	1.01
Number of oocytes spline	Spline with three knots. Not directly interpretable.				
1		1.148	0.173 to 2.130	1340	1.00
2		-1.573	-2.634 to -0.494	1511	1.00
3		1.369	-0.087 to 2.809	1989	1.00
Mean fragmentation of embryos selected for transfer	Standardised, linear variable	-0.152	-0.233 to -0.071	3961	1.00

S Table 26: Characteristics of the double embryo transfer submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to log(odds ratios). Odds ratios indicate the relative change in the odds of having two rather than one embryos being transferred.

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Embryo implantation submodel	Probit regression submodel				
Intercept	Mean Z score, after averaging over other variables	0.112	-1.139 to 1.379	2610	1.00
Long Agonist (vs Antagonist)	Dummy indicator variable	-0.041	-0.183 to 0.103	10500	1.00
Total dose spline	Spline with knots at quintiles. Not directly interpretable.				
1		-1.227	-2.521 to 0.083	2981	1.00
2		-0.431	-1.402 to 0.500	2838	1.00
3		-0.900	-1.854 to 0.064	2646	1.00
4		-0.779	-1.726 to 0.159	2659	1.00
5		-1.335	-2.405 to -0.241	3839	1.00
6		-1.498	-2.861 to -0.146	4107	1.00
7		-0.939	-2.378 to 0.452	4461	1.00
HMG (vs FSH)	Binary indicator variable	0.095	-0.047 to 0.237	10500	1.00
Age spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.199	-1.596 to 1.198	3119	1.00
2		0.408	-0.434 to 1.251	3366	1.00
3		-0.098	-1.061 to 0.875	2801	1.00
4		0.380	-0.516 to 1.304	2862	1.00
5		-0.350	-1.383 to 0.669	2992	1.00
6		0.521	-0.523 to 1.587	2919	1.00
7		-0.923	-2.165 to 0.312	3543	1.00
Number of cleaved embryos spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.180	-3.251 to 2.862	7743	1.00
2		0.918	-0.574 to 2.404	3076	1.00
3		0.665	0.158 to 1.128	733	1.00
4		0.591	0.168 to 1.004	336	1.01
5		0.854	0.043 to 1.649	337	1.01
6		0.924	-0.355 to 2.183	427	1.01
7		-0.393	-1.965 to 1.149	629	1.00
Number of oocytes spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.016	-1.070 to 1.054	562	1.00
2		0.143	-1.110 to 1.401	808	1.00
3		-0.067	-1.686 to 1.598	899	1.00
4		0.122	0.042 to 0.201	2250	1.00
5		-0.178	-0.614 to 0.287	327	1.00
6		-0.111	-0.490 to 0.272	10500	1.00
7		0.018	-0.210 to 0.245	10500	1.00
Mean fragmentation of embryos selected for transfer	Standardised, linear term	-0.177	-0.401 to 0.038	10500	1.00

S Table 27: Characteristics of the embryo implantation submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to change in ‘Z scores’, which correspond to changes in probability that one or more of the transferred embryos will implant as each covariate varies. Continues in S Table 28:

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Embryo implantation submodel, continued	Probit regression submodel				
Double embryo transfer (vs single)	Binary indicator variable	-0.582	-0.841 to -0.332	10500	1.00
Practitioner performing transfer (vs 12)	Categorical variable w 13 levels. 12 used as reference category due to highest frequency.				
1 (fewer than 20 procedures)		0.040	-0.176 to 0.262	10500	1.00
2		0.184	-0.029 to 0.388	10500	1.00
3		0.006	-0.245 to 0.247	10500	1.00
4		0.130	-0.079 to 0.341	10500	1.00
5		0.175	-0.226 to 0.590	10500	1.00
6		0.210	-0.036 to 0.459	10500	1.00
7		0.015	-0.151 to 0.177	6651	1.00
8		0.029	-0.178 to 0.234	10500	1.00
9		-0.035	-0.212 to 0.149	3443	1.00
10		-0.095	-0.594 to 0.398	10500	1.00
11		-0.444	-1.020 to 0.108	10500	1.00
13		0.020	-0.174 to 0.210	4983	1.00
Diagnosis	Seven binary indicators				
Tubal disease		-0.043	-0.211 to 0.127	2165	1.00
Ovulation Failure		-0.119	-0.397 to 0.162	10500	1.00
Uterine problem		0.010	-0.172 to 0.195	4048	1.00
Unexplained		-1.711	-2.787 to -0.683	2136	1.00
Male factor		0.423	-1.039 to 1.949	2619	1.00
Endometriosis		0.230	-0.648 to 1.158	2674	1.00
Anovulation		0.430	-0.587 to 1.494	2371	1.00

S Table 28: Continuation of S Table 27: Characteristics of the embryo implantation submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to change in ‘Z scores’, which correspond to changes in probability that one or more of the transferred embryos will implant as each covariate varies. Continues in S Table 28:

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Live birth event submodel	Probit regression submodel				
Intercept	Mean Z score, after averaging over other variables	1.826	0.203 to 3.444	1681	1.00
Long Agonist (vs Antagonist)	Dummy indicator variable	-0.058	-0.286 to 0.174	10500	1.00
Total dose spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.893	-2.712 to 0.794	3500	1.00
2		-0.062	-1.279 to 1.111	3087	1.00
3		-0.986	-2.198 to 0.130	2556	1.00
4		-0.589	-1.795 to 0.523	2885	1.00
5		-1.007	-2.454 to 0.402	3429	1.00
6		-0.548	-2.445 to 1.362	3828	1.00
7		-0.686	-2.320 to 1.027	4157	1.00
HMG (vs FSH)	Binary indicator variable	-0.098	-0.333 to 0.133	10500	1.00
Age spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.050	-1.825 to 1.657	3915	1.00
2		-0.189	-1.314 to 0.945	4168	1.00
3		0.329	-0.875 to 1.488	3307	1.00
4		-0.283	-1.416 to 0.798	3276	1.00
5		0.564	-0.707 to 1.831	3615	1.00
6		-1.111	-2.663 to 0.332	3362	1.00
7		0.777	-1.186 to 2.908	6192	1.00
Number of cleaved embryos spline	Spline with knots at quintiles. Not directly interpretable.				
1		0.997	-1.279 to 3.285	5716	1.00
2		-0.598	-1.737 to 0.549	1257	1.00
3		0.345	-0.351 to 1.055	500	1.00
4		0.265	-0.373 to 0.932	284	1.01
5		-0.530	-1.558 to 0.565	276	1.01
6		1.120	-0.489 to 2.746	355	1.01
7		-1.321	-3.273 to 0.542	518	1.01

S Table 29: Characteristics of the live birth event submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to change in ‘Z scores’, which correspond to changes in probability that any of the implanted embryos will be sustained, resulting in a live birth event, as each covariate varies. Continued in S Table 30

Variables	Comments	Estimate	95% CI	Effective sample size	Gelman-Rubin convergence statistic
Live birth event submodel, continued	Probit regression submodel				
Number of oocytes spline	Spline with knots at quintiles. Not directly interpretable.				
1		-0.866	-2.201 to 0.461	4824	1.00
2		-0.158	-0.983 to 0.657	2769	1.00
3		-0.672	-1.636 to 0.240	1538	1.00
4		-0.094	-1.029 to 0.863	802	1.00
5		-0.987	-2.320 to 0.315	705	1.00
6		0.237	-1.737 to 2.162	901	1.00
7		1.802	-0.814 to 4.815	3774	1.00
Mean fragmentation of embryos selected for transfer	Standardised, linear term	-0.127	-0.260 to 0.002	1371	1.00
Double embryo transfer (vs single)	Binary indicator variable	0.101	-0.495 to 0.667	184	1.02
Practitioner performing transfer (vs 12)	Categorical variable w 13 levels. 12 used as reference category due to highest frequency.				
1 (fewer than 20 procedures)		0.039	-0.599 to 0.735	10500	1.00
2		0.319	-0.097 to 0.754	5474	1.00
3		-0.352	-0.721 to 0.013	10500	1.00
4		-0.025	-0.550 to 0.530	2640	1.00
5		-0.088	-0.423 to 0.273	6930	1.00
6		-0.195	-0.500 to 0.123	10500	1.00
7		0.110	-0.290 to 0.525	6944	1.00
8		-0.227	-0.558 to 0.103	6742	1.00
9		-0.399	-0.965 to 0.179	10500	1.00
10		-0.038	-0.402 to 0.336	10500	1.00
11		-0.323	-0.579 to -0.070	6169	1.00
13		0.091	-0.257 to 0.438	6460	1.00
Diagnosis	Seven binary indicators				
Tubal disease		-0.118	-0.398 to 0.163	3139	1.00
Ovulation Failure		-0.447	-1.213 to 0.339	10500	1.00
Uterine problem		-0.565	-1.516 to 0.385	10500	1.00
Unexplained		-0.062	-0.366 to 0.233	5972	1.00
Male factor		0.023	-0.242 to 0.285	4081	1.00
Endometriosis		-0.090	-0.566 to 0.385	10500	1.00
Anovulation		0.008	-0.286 to 0.300	4800	1.00

S Table 30: Continuation of S Table 29. Characteristics of the live birth event submodel, including details of covariates, parameter estimates and convergence diagnostics. Estimates (95% CIs) in this submodel correspond to change in ‘Z scores’, which correspond to changes in probability that any of the implanted embryos will be sustained, resulting in a live birth event, as each covariate varies.

11.7.3. **Additional Supplementary Material for Chapter 11.**

Supplementary Material 1: Mathematical presentation of the model.

Following the notation of Journal Article 5 (Chapter 10):

Dose submodel

For patient *j*, we model the logarithm of the total dose of gonadotropins (y_j^T), using linear regression:

$$y_j^T = \boldsymbol{X}_j^T \boldsymbol{\beta}^T + z_j^T$$

where $\boldsymbol{X}_j^T$ is a row-vector of covariates, $\boldsymbol{\beta}^T$ is a vector of regression parameters, and z_j^T is the model residual.

Number of oocytes submodel

For patient *j,* we model the number of oocytes (y_j^O) using Poisson regression, with $y_j^O \sim Poisson(\lambda_j^o)$:

$$\log(\lambda_j^o) = \boldsymbol{X}_j^o \boldsymbol{\beta}^o + \boldsymbol{U}_j^o \boldsymbol{\delta}^o + z_j^o$$

Where λ_j^o is the rate parameter, $\boldsymbol{X}_j^o$ is a row-vector of covariates, $\boldsymbol{\beta}^o$ is a vector of regression parameters and z_j^o is a patient-specific latent variable. $\boldsymbol{U}_j^o$ is a row-vector of 'outcome-covariates' corresponding to upstream response variables (in this case, a spline representation of total dose) and $\boldsymbol{\delta}^o$ is a corresponding vector of regression parameters**..**

Fertilisation submodel

We model the number of embryos obtained when oocytes are mixed with sperm y_j^M in terms of its rate parameter λ_j^M, again using a Poisson submodel:

$$\log(\lambda_j^M) = \log(y_j^O) + \boldsymbol{X}_j^M \boldsymbol{\beta}^M + \boldsymbol{U}_j^M \boldsymbol{\delta}^M + z_j^M$$

where $\boldsymbol{X}_j^M$,$\boldsymbol{\beta}^M$ and z_j^M are analogous to the corresponding terms in the stimulation model. We now include an offset term $\log(y_j^O)$ corresponding to the logarithm of the number of oocytes obtained in the linear predictor. $\boldsymbol{U}_j^M$ is a row-vector of 'outcome-covariates' corresponding to upstream response variables (in this case, a spline representation of total dose) and $\boldsymbol{\delta}^M$ is a corresponding vector of regression parameters**..**

Embryo fragmentation submodel

Degree of fragmentation (y^F) is an ordinal 1 to 4 grading scale. We model this using cumulative logit regression. For embryo *i* (where *i* = 1,2,…,n_j) and patient *j* we have, for *k* = 1,2,3:

$$\text{logit}(\gamma_{kij}^F) = \alpha_k^F - \boldsymbol{X}_{ij}^F\boldsymbol{\beta}^F + \boldsymbol{U}_j^F\boldsymbol{\delta}^F - z_j^F$$

where $\boldsymbol{X}_{ij}^F$ is a row-vector of covariates, $\boldsymbol{\beta}_k^F$ is a vector of regression coefficients and z_j^F is a random effect. γ_{kij}^F is a cumulative probability of embryo *i* in patient *j* having a grade of *k* or lower for fragmentation degree and α_k^F is a threshold parameter, corresponding to the log-odds of the embryo having grade *k* or lower. $\boldsymbol{U}_j^F$ is a row-vector of 'outcome-covariates' corresponding to upstream response variables (in this case, spline representations of total dose and number of oocytes) and $\boldsymbol{\delta}^F$ is a corresponding vector of regression parameters.

Double embryo transfer submodel

We model double embryo transfer using probit regression. Let $y_j^D = 1$ or 0 if patient *j* does or does not have DET, respectively. We define y_j^{D*} as a latent continuous variable underlying the binary y_j^D, such that:

$$y_j^D = \begin{cases} 1\ if y_j^{D*} \ \geq 0 \\ 0\ if\ y_j^{D*} \ < 0 \end{cases}$$

A linear regression submodel for the latent y_j^{D*} is then used to estimate covariate effects:

$$y_j^{D*} = \boldsymbol{X}_j^D\boldsymbol{\beta}^D + \boldsymbol{U}_j^D\boldsymbol{\delta}^D\ z_j^D$$

$$z_j^D \sim N(0,1)$$

where $\boldsymbol{X}_j^D$ is a row-vector of patient-level covariates and $\boldsymbol{\beta}^D$ is a vector of regression coefficients. $\boldsymbol{U}_j^D$ is a row-vector of 'outcome-covariates' corresponding to upstream response variables (in this case, spline representations of total dose, number of oocytes and number of cleaved embryos, and mean fragmentation of transferred embryos) and $\boldsymbol{\delta}^D$ is a corresponding vector of regression parameters.

Implantation and Live birth event submodels

As for DET, we use probit regression to model implantation and live birth event (LBE). $y_j^I = 1$ or 0 and $y_j^L = 1$ or 0 if there is/is not an implantation and LBE, respectively, with underlying latent variables y_j^{I*} and y_j^{L*} , row vectors of patient-level covariates $\boldsymbol{X}_j^I$ and $\boldsymbol{X}_j^L$ and vectors of regression coefficients $\boldsymbol{\beta}^I$ and $\boldsymbol{\beta}^L$. Additionally, these submodels contain row-vectors of outcome-covariates $\boldsymbol{U}_j^I$ and $\boldsymbol{U}_j^{\boldsymbol{L}}$ (containing spline representations of total dose, number of oocytes, and number of cleaved embryos, in addition to mean fragmentation of transferred embryos and an indicator denoting double embryo transfer), and corresponding vectors of regression parameters $\boldsymbol{\delta}^I$ and $\boldsymbol{\delta}^L$**.** The error terms z_j^I and z_j^L have variance 1.

Latent variable distribution

We specify a multivariate Normal distribution for the latent variables to connect the submodels:

$$\begin{bmatrix} z_j^T \\ z_j^O \\ z_j^M \\ z_j^F \\ z_j^D \\ z_j^I \\ z_j^L \end{bmatrix} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$$

where we estimate the elements of $\boldsymbol{\Sigma}$ together with the rest of the model.

Further details can be found at : http://www.biorxiv.org/content/early/2017/08/10/173534

Supplemental Material 2: Stan code to fit the model

```
data {
  int <lower=0> Nstart;//number of cycles started

  int <lower=0> Nmix; //number of cycles with eggs mixed with
sperm
  int <lower=0> Ntrans; //number of cycles reaching transfer
  int <lower=0> Nembryo; //number of embryos
  int <lower=0> Nimp; // number of cycles achieving embryo
implantation
```

```
  int <lower=0> Pstart; //number of regression params in starting
model
  int <lower=0> Pmix; //number of regression params in
fertilisation model
  int <lower=0> Pdet; //number of params in DET model
  int<lower=0> Pimp; //number of params in implantation model
  int <lower=0> Plbe; //number of params in LBE model
  int <lower=0> Pembryo; //number of params in embryo models
  int<lower=0> Pdose; //number of regression params in dose model


  int<lower=0> Mtrans; //number with no trans
  int<lower=0> Mimp; //number with no implantation

  int <lower=0> cid[Nembryo]; //cycle id number at embryo level
  int <lower=0> startid[Nstart]; //cycle id number, cycle-level,
all cycles started
  int <lower=0> mixid[Nmix]; //cycle id numbers, subsetted to
cycles with cleavage
  int <lower=0> transid[Ntrans]; //cycle id numbers, subsetted to
cycles with transfers
  int <lower=0> impid[Nimp]; //cycle id numbers, subsetted to
cycles with implantation
  int <lower=0> mtransid[Mtrans];
  int <lower=0> mimpid[Mimp];


  matrix[Nembryo, Pembryo] Xembryo; //design matrix for embryo
submodels
  matrix[Nstart, Pstart] Xstart;
  matrix[Nmix, Pmix] Xmix;
  matrix[Ntrans, Pdet] Xdet;
  matrix[Nimp, Plbe] Xlbe;//lbe covariates - only in those who had
implantation
  matrix[Nstart, Pdose] Xdose; //design matrix for dose submodel
  matrix[Ntrans, Pimp] Ximp; //implantation covariates, in those
who had transfer

  //Response variables and offset for fert model

  vector<lower=1>[Nmix] mTotEgg;
  int<lower=0> ncleave[Nmix];
  int<lower = 0> TotEgg[Nstart];
  row_vector<lower=-1, upper = 1>[Ntrans] detsign; //det = 0 -> -
1, det = 1 -> 1
  row_vector<lower=-1, upper = 1>[Ntrans] impsign; //imp = 0 -> -
1, imp = 1 -> 1
  row_vector<lower=-1, upper = 1>[Nimp] lbesign;


  int<lower=0,upper=1> cumfrag1[Nembryo];
  int<lower=0,upper=1> cumfrag2[Nembryo];
  int<lower=0,upper=1> cumfrag3[Nembryo];
  vector[Nstart]ltotdose;
```

```
}


parameters{
  vector[Pdose] betad;
  vector[Pstart] betao;
  vector[Pmix] betam;
  vector[3] alphaof;
  vector[Pembryo] betaof;
  vector[Pdet] betadet;
  vector[Pimp] betaimp;
  vector[Plbe] betalbe;

  matrix[3,Nstart] Z24;//2nd to fourth random effects, of seven,
all at cycle level
  cholesky_factor_corr[7] L;//corr matrix for lv 2 random
effects/latent variables
  row_vector<lower=0>[Ntrans] obs_abs_detstar;//latent var for
observed DET
  row_vector[Mtrans] unobs_detstar;
  row_vector<lower=0>[Ntrans] obs_abs_impstar;//latent var for
observed implantation outcomes
  row_vector[Mtrans] unobs_impstar;
  row_vector<lower=0>[Nimp] obs_abs_lbestar;//latent var for
observed birth outcomes
  row_vector[Mimp] unobs_lbestar;//latent var for unobs birth
outcomes-


  vector<lower=0>[4] theta14;//four variance parameters of seven-
put prior onto this vector


}







  transformed parameters{


  matrix[7,Nstart] Z;
  vector[7] theta;
  vector[Nmix] lTotEgg;
  vector[7] mu;
  vector[Nstart] XBstart;
  vector[Nmix] XBmix;
  vector[Nembryo] XBfrag;
  vector[Ntrans] XBdet;
  vector[Nimp] XBlbe;
  vector[Nstart] XBdose;
```

```
vector[Ntrans] XBimp;



XBdose<-Xdose*betad;//linear predictor for dose
XBstart<-Xstart*betao; //linear predictor for oocytes
XBmix<- Xmix*betam; //linear predictor for fert
XBfrag<-Xembryo*betaof; //linear predictor for frag
XBdet<-Xdet*betadet; //linear predictor for DET
XBimp<-Ximp*betaimp;//linear predictor for implantation
XBlbe<-Xlbe*betalbe; //linear predictor for LBE





lTotEgg <-log(mTotEgg);

for (n in 1:Nstart)//Z1 is the residual from the dose model
Z[1, startid[n]]<- ltotdose[n]-(XBdose[n]);

for (p in 1:Ntrans){
  Z[5,transid[p]]<- detsign[p] * obs_abs_detstar[p] - XBdet[p];
}

for (u in 1:Mtrans){
  Z[5,mtransid[u]]<-unobs_detstar[u];
}

for (p in 1:Ntrans){
  Z[6,transid[p]]<- impsign[p] * obs_abs_impstar[p] - XBimp[p];
}

for (u in 1:Mtrans){
  Z[6,mtransid[u]]<-unobs_impstar[u];
}





for (p in 1:Nimp){
  Z[7,impid[p]]<- lbesign[p] * obs_abs_lbestar[p]-XBlbe[p];
}

for (u in 1:Mimp){
  Z[7,mimpid[u]]<-unobs_lbestar[u];
}

for (k in 2:4){Z[k]<-Z24[k-1];}

theta[5]<-1.0;
theta[6]<-1.0;
theta[7]<-1.0;
  for(i in 1:4){theta[i]<-theta14[i];}
  for (k in 1:7)
  mu[k]<-0;


}



model{


  betad~normal(0,10);
  betao~normal(0,10);
  betam~normal(0,10);
  alphaof~normal(0,10);
  betaof~normal(0,10);
  betadet~normal(0,2);
  betaimp~normal(0,2);
  betalbe~normal(0,2);
  theta14~cauchy(0,2.5); //half cauchy - see 6.12
  L~lkj_corr_cholesky(1);



  //dose model
   ltotdose~normal(XBdose, theta[1]);



   //stimulation model
  for (i in 1:Nstart){
  TotEgg[i]~poisson_log(XBstart[i]+ Z[2,startid[i]]);}


  //transfer models - residuals calculated in transformed
parameters block
  for (i in 1:Nstart){

    Z[,startid[i]]~multi_normal_cholesky(mu,
diag_pre_multiply(theta,L));}




//embryo models

 for (i in 1:Nembryo){



    //fragmentation degree
    cumfrag1[i]~bernoulli_logit(alphaof[1]-
(XBfrag[i]+Z[4,cid[i]]));

    cumfrag2[i]~bernoulli_logit(alphaof[2]-
(XBfrag[i]+Z[4,cid[i]]));

    cumfrag3[i]~bernoulli_logit(alphaof[3]-
(XBfrag[i]+Z[4,cid[i]]));}




//fertilisation model

for (i in 1:Nmix) {

  ncleave[i]~poisson_log(lTotEgg[i]+XBmix[i]+ Z[3,mixid[i]]);


}




}




generated quantities{
    matrix[7,7] Eta;
    matrix[7,7] Theta;
    Eta<-multiply_lower_tri_self_transpose(L);
    Theta<-quad_form_diag(Eta, theta);

  }
```

IV. Conclusions

Chapter 12. Discussion of the thesis

12.1 How should we measure multistage IVF outcomes?

It became apparent almost immediately during a review of the literature (Chapter 1) that the goal of identifying a single outcome measure for IVF was fanciful. Different outcome measures are more or less useful for different purposes, and for different audiences. This somewhat blithe statement belies the complexity of the problem however. Having reviewed both the range of expert opinions (Chapter 1) and the outcomes in use in several settings (Journal Articles 1, 2 and 3), as well as having solicited the opinions of many IVF 'survivors' (to borrow a term in use in online patient groups), we have identified a wide array of elements that must be considered when selecting the right outcome measure for a given purpose. The problem of what to report arises due to the plethora of options on offer. This in turn results from the multistage nature of treatment, which produces a sequence of response variables (potential numerators) and opportunities for left truncation (hence, potential denominators). There is neither consensus nor consistency in how numerators are measured. In Journal Article 2 (Chapter 4), we identified 361 numerators in use in IVF RCTs, after combining similar variables (and hence understating the actual variation in reporting). We identified 87 distinct denominators (again, after combining similar items), resulting in 815 distinct combinations. Even on IVF clinic websites, where reported outcomes should be restricted to measures of clinical relevance, we found considerable variation in reporting (Journal Article 1, Chapter 3). We identified 54 outcome measures in use, including 33 different ways of reporting pregnancy and 9 different ways of reporting live birth. The scope for variety was expanded by modifiable reporting filters, defined by date ranges, or patient and treatment characteristics.

The implications are manifold. In the context of RCTs, there is scope for selective reporting, including flexibility in adopting data-driven outcome definitions from a menu of potential measures, all of which are consistent with the trial protocol. This does not require conscious cheating on behalf of the investigators. Rather, subtle decisions made at the point of analysis may bias a study's results even when made in good faith (Gelman and Loken, 2013, Simmons, et al., 2011). It has recently been argued that core (mandatory, standardised) outcome sets reduce the scope for p-hacking of this sort (van't Hooft and Khan, 2017) and there is *prime facie* some credibility to this claim. The adoption of core outcome sets in IVF trials might also go some way to resolving a second concern relating to outcome heterogeneity; namely, the hindrance this presents for systematic review and meta-analysis.

In the context of IVF clinic websites, financial conflicts of interest in combination with the extensive menu of reporting options exacerbates concerns about selective reporting. Potential patients may be misled on the basis of cherry-picked results, or else by superficially equivalent but substantively different measures of success used by clinics. As for outcome reporting in RCTs, standardised measures offer one solution to the problem. National reporting schemes such as that curated by HFEA in the UK or SART in the US present standardised measures of clinic performance. This reduces selective reporting by clinics, but does not necessarily remove gaming entirely. Clinics remain free to select and treat their patients so as to maximise their ranking according to the national performance indicator. These behaviours may not guarantee the best interests of patients. Where national reporting schemes are in place, a second concern relates to the particular measure or measures chosen as the universal standard. In Journal Article 3 (Chapter 5), we reviewed the current reporting practices of HFEA and SART, and raised concerns over the new standards introduced recently by the former. In particular, we objected to the implicit exclusion of patients with failed ovarian stimulation from the measures 'live birth event per embryo transferred' and 'cumulative live birth per egg collection'. We also objected to the former on the grounds of poor face validity, unit of analysis error, and the fact that it removed relevant variation in clinic policies, which is a fatal limitation for a performance measure (Bird, et al., 2005). It appears that HFEA solicited at most very limited involvement from statisticians when consulting on the choice of measure.

And therein lies a recurring theme in discussion and deployment of IVF outcome measures. While the debate tends to acknowledge the complexity of balancing competing perspectives of clinicians and patients, and even of those with financial incentives, statistical considerations do not regularly or prominently feature. One attitude towards the role of statistics in determining outcomes was encapsulated in a response to talk given by the lead author on the topic of outcome reporting on IVF clinic websites, at the

Cochrane Gynaecology and Fertility 20 year anniversary meeting, in April of 2016. The comment was made that IVF clinic websites were 'not scientific papers', with the implication being that the statistical arguments being advanced did not matter. While it is true that stakeholders should decide exactly what content is important and should be conveyed by outcome measures, it is the role of statistics to assess whether a measure has any content at all (for example, that there is no unit of analysis error being made) and, if so, to compare that content against whatever the measure was intended to convey. A measure that excludes patients with cancelled stimulation cycles from the denominator, for example, conditions on successful ovarian stimulation and therefore offers neither prognostic utility nor information about variation in stimulation performance, which is a key determinant of downstream success.

At least we can agree with the commenter that statistical considerations matter in relation to outcome measurements in the domain of IVF research. We've already commented on the plethora of measures in use in IVF RCTs. This might not be a sign of methodological malaise *per se*; it could be, in principle, that each measure we encountered was selected with the intention of answering a particular research question, and was appropriate for that purpose. This doesn't appear to be the case, however. During our review of RCTs, we looked at the denominators used to report the outcome 'live birth' and its variants. Over half of them used a denominator that didn't represent the full randomised cohort (we did not include trials with small numbers of exclusions on grounds of non-adherence here, since this is a separate issue). We saw examples where the (valid) measurement of treatment effect on live birth was precluded by the trial design. Trials where each patient's oocytes were randomly divided between two interventions constitute one example. Resulting embryos from one of the arms (whichever were 'best') were transferred to the patient, and the live birth outcome was attributed to whichever group they came from. This does not represent a randomised comparison. Where dysfunctional outcome measures are being used as the basis for inference, the clinical conclusions arrived at must be subject to doubt.

Arguably perhaps, throughout Part II of the thesis, we have spent less time addressing the question of how to measure IVF and more time addressing the matter of how not to. Given that so much of what is done could be classified as examples of the latter, we feel that this is itself an important contribution to the topic. Drawing attention to methodological limitations and thereby promoting scepticism towards reported IVF outcomes is something that appears to be urgently required. A recent review of treatments offered by IVF clinics in the UK found many of them lack good quality evidence of clinical benefit (Heneghan, et al., 2016). A response by UK practitioners served to underline concerns that much practice is not driven by sound scientific reasoning. They responded that, in their view, there was good quality evidence, and due to the fact that IVF is complicated, the effectiveness of interventions can often not be established by RCTs (Balen, et al., 2017). These are the same kinds of arguments frequently advanced by practitioners of alternative medicine. Given that research methodology does not appear to be well understood by many who practice and research IVF, the need for antagonistic statistical voices appears to be acute. This is particularly true in light of the vulnerable patient population and costs involved.

That said, our contribution has not been entirely negative. In relation to direct advertising of IVF to patients by clinics, we have argued for an outright ban on the basis of our findings in Journal Article 1 (Chapter 3). Instead, we argue, reporting of success rates should be restricted to standardised public reporting schemes. We have also argued that measures used for this purpose should generally include all women who begin treatment in the denominator (Journal Article 3, Chapter 5). Moreover, we have argued that a set, rather than a single measure, should be presented, in order to capture different aspects of the clinic's performance and to put greater emphasis on safety. While we presented a possible set of outcome measures for this purpose in Journal Article 3, it seems appropriate to disclose that this had not been our intention when submitting that manuscript for publication. It was at the insistence of a peer reviewer that a particular set of measures was included. While we included our preferred set, we believe that this should form the basis of further discussion rather than be taken as the last word on the topic.

Our proposed set of measures included the incidence of ovarian hyperstimulation syndrome (OHSS). One theme that became apparent during the thesis was the disconnect between IVF research and public reporting in relation to safety concerns over ovarian stimulation. There is currently substantial research interest in the idea of tailoring ovarian stimulation to the patient so as to avoid excessive responses, which are associated with OHSS. This was the motivating idea behind the analysis presented in Journal Article 4 (Chapter 9). Recent research includes the validation of predictive markers (Broer, et al., 2011, Broer, et al., 2013a, Broer, et al., 2013b, Rustamov, et al., 2011, Rustamov, et al., 2012, Rustamov, et al., 2014) and both development (La Marca, et al., 2013, La Marca, et al., 2012) and testing (Allegra, et al., 2017, Nyboe Andersen, et al., 2017, Popovic-Todorovic, et al., 2003) of dose-selection algorithms designed to reduce variation in ovarian response. However, we found that incidence of OHSS was not reported by any clinic website in our review. At the time of writing, OHSS rates also do not feature on the patient-directed websites of either HFEA or SART. While regulators have done well to address safety concerns associated with twin births from IVF, it is concerning if prospective patients are not being sufficiently warned of the risks of ovarian stimulation. If potential harms of treatment are not clearly communicated, then prospective patients are instead left to base the decision of whether and where to be treated on some combination of success rates and invariably glowing patient testimonials. Based on a feeling that the downsides of IVF are often downplayed by those selling the treatments, a new and ongoing online campaign set up by IVF survivors is seeking to add some realism to the otherwise selective narratives seen by women researching treatments online, by encouraging women to share their unfiltered experiences of IVF (Repro Tech Truths, 2017).

Clearly, the matters of IVF clinic outcome reporting and direct advertising to patients have not been settled, but patient-led initiatives such as reprotechtruths.org, as well as ongoing research into reporting standards (at the time of writing, we are aware of ongoing reviews in Latin America and Europe, in addition to our own UK review) are setting the stage for greater transparency. Following the publication of our website review, HFEA have told us that they will put the issue of outcome reporting in the UK on their agenda. It remains to see if this happens, and if so what their solution will be.

At the time of writing, we have not made suggestions regarding outcome measures for IVF clinical trials. This is partially because we believe that the appropriate outcome measure for a trial depends on the particular research question at hand. There is a potential conflict between this view and the IMPRINT (Improving the Reporting of

Infertility Trials) statement, which encourages live birth to be reported for all trials (Legro, et al., 2014). For the within-patient designs described above, where batches of oocytes are randomly allocated to different interventions, this would require an inappropriate outcome measure to be reported. Similarly, ongoing efforts as part of the CROWN (Core Outcomes in Women's Health) initiative appear to ignore the possibility that valid reporting of some outcome measures might be precluded by some kinds of design (Khan, 2014). While the goal of aiming towards commensurability of outcome measures for systematic review and comparison between treatments is admirable, there is no value in a meta-analysis of the outcome 'live birth event per whichever of two half-batches of oocytes randomised produced the best embryos'.

The problem largely appears to be attributable to the fact that studies within CROWN do not clearly delineate between effectiveness trials (concerned with the question of whether introducing a treatment actually leads to improved patient outcomes) and efficacy trials (earlier phase trials which test whether or not an intervention has the intended effect on the IVF process). This is in contrast, for example, to COMET (Core Outcome Measures in Effectiveness Trials) (Williamson, et al., 2011). Although it does not permit an assessment of clinical effectiveness, the within-patient design is excellent for the purpose of testing whether an intervention produces superior embryos (or, for example, higher fertilisation rates). This is because it eliminates within-patient sources of variation, and requires smaller sample sizes compared to a between-patient design. Accordingly, if the intervention does not work as intended, this can be established with a smaller cohort. On the other hand, if the treatment shows efficacy in principle, it can then be evaluated as part of a practical treatment strategy in larger studies. If we were to follow the recommendations of IMPRINT and CROWN, and to insist that (valid) patient-centred measures must be included in all trials, this breed of within-patient design would not be admissible[5].

If women's health journals are committed to making sure all trials adhere to CROWN then, we might be facing a future where many trials are mandated to report unsuitable

[5] We do not enter into the debate around whether or not a second type of within-patient design, the crossover trial, is suitable for IVF studies. See Vail and Gardener 2003, Makubate and Senn, 2010 and McDonnell et al., 2004.

outcome measures or else to adhere to study designs which permit clinical effectiveness (and safety) endpoints to be measured. In fact, the idea of core study designs has already been introduced (Myatt, et al., 2014). It is unclear how to react to this as methodologists. On the one hand, it is difficult to see how all possible research questions can be anticipated and accommodated by standardised trial designs. On the other hand, any applied methodologist can attest to the recurrence of poor methodological design features in the clinical literature. To paraphrase Doug Altman (and to oversimplify an issue with multifactorial causes), this happens because many of the people doing research are not competent to do so (Altman, 1994). To the extent that core designs are based on methodological principles, a totalitarian publication policy, in which studies are accepted on the condition that they adhere to standardised design principles, might lead to a net improvement in the quality of the literature. Cochrane could be taken as a model. Authors produce systematic reviews which are published on the condition that they achieve the standards set out by the Cochrane Handbook, which is primarily authored by methodologists (Higgins and Green, 2011). Flexibility is permitted where departures from the guidelines can be justified for the problem at hand. The Cochrane model is not uncontroversial however. As Stephen Senn has commented on Twitter: “The aviation equivalent of the Cochrane Collaboration would be an organization writing manuals to instruct kids to fly jumbo jets”.

As we discussed in the context of public-facing outcome measures, the key point here is that statistical and methodological considerations must be taken into account when choosing outcome measures for IVF trials. Reassuringly, the fledgling COMMIT project (Core Outcome Measures for Infertility Trials) will include statisticians amongst its steering group. The project will use the review of trials presented in the thesis (Journal Article 2, Chapter 4) as a starting point.

One additional project we intend to complete in the imminent future is the production of a list of *statistically valid* outcome measures for IVF RCTs. The measures will avoid the common problems we have identified as stemming from the multistage treatment structure, such as improper subgrouping and unit of analysis errors. This will not be restricted to clinically relevant endpoints, but, using Journal Article 2 as a basis, will include outcome measures suitable for answering research questions at each stage of the

IVF cycle. We will include explanations as to what makes these outcome measures preferable over common alternatives. We will emphasise the importance of the choice of denominator for preserving the advantages proffered by randomisation. We would then hope that the core outcomes chosen as part of the COMMIT process would comprise the intersection of the sets of clinically relevant and of statistically valid outcomes measures, while emphasising the fact that core outcomes might not be suitable for efficacy and mechanism trials. This would be a valuable resource to the IVF research community, and one that affords a greater degree of personal responsibility than does the strategy of mandating conformity in every aspect of design.

Most of the issues we have discussed here had not been anticipated at the outset. The primary motivation for reviewing IVF outcome measures was to inform the development of statistical methods for multistage IVF data. By investigating which events in the IVF cycle were commonly reported, how they were measured, and the questions they were used to answer, we were able to flesh out our understanding of the underlying processes which we sought to represent through modelling. During the process of conducting these review exercises however, it became apparent that there were important statistical issues that needed to be spotlighted. Journal Articles 1, 2 and 3 represent our attempts to do so. These simple review and discussion papers might ultimately lead to greater patient benefit than the complex methodological work, which we turn our attention to next.

12.2 How should we model multistage IVF data?

Relatively little work has been done on the analysis of multistage IVF data up to now. Two proposals to use discrete time to event methods for this purpose have been advanced (Maity, et al., 2014, Penman, et al., 2007). We have discussed the limitations of these approaches (wasteful outcome dichotomisation, inability to handle different covariates at different treatment stages) throughout the thesis. These methods are useful for the purpose of assessing association between baseline covariates and the probability of success or failure at a given stage of the cycle. They are less suitable for the purpose of answering mechanistic research questions about relationships between interventions and events at different stages of the cycle.

Accordingly, we identified, adapted and developed methods in use in other fields, including econometrics, education research, and developmental toxicology, for the purpose of analysing multistage treatment data with mixed, multilevel responses. After investigating the matter of how to model stimulation response and embryo quality (Chapter 6.), we proceeded to jointly model response variables by positing an underlying multivariate Gaussian latent structure (Chapter 7.). The initial challenge was to establish suitable representations of each response variable, so that they could be connected to the other responses by way of this latent structure. The work of Gueorguieva and colleagues (Gueorguieva, 2001, Gueorguieva and Agresti, 2001, Gueorguieva and Sanacora, 2006) describing joint models for different combinations of response types was useful in this regard, as was Goldstein and colleagues' framework, which accommodated mixed responses defined at different levels of a multilevel data structure (2009). The measures of latent association yielded by this approach were not obviously interpretable however. They represent correlation coefficients and remain unadjusted for other response variables in the model. As such, they do not provide clear effect estimates on the scales of the responses.

An obvious answer to this problem was to allow response variables in the model to enter as 'outcome-covariates' in the submodels relating to the downstream stages of the cycle. It would not be appropriate to model the cycle as a set of unconnected 'outcome-regression' submodels however, since to do so would be to assume that there was no unmeasured confounding between the model response variables. This might still be superior to the correlated latent variable approach, since we can at least adjust for upstream response variables as covariates in each submodel. However, the likelihood of residual confounding due to unmeasured variables and measurement error in the included covariates means that univariate regression modelling is unlikely to be sufficient for valid inference. Consequently, we accommodated unmeasured confounding by maintaining the underlying multivariate latent Gaussian structure at the same time as including response variables as model covariates. This explicitly models the correlation between response variables resulting from unmeasured confounding (so-called *endogeneity*). We refer to this, our preferred method for the analysis of multistage IVF data, as the endogenous response model. Turning a correlated latent variable model into an endogenous response model is easy, since one just has to include upstream responses in the appropriate covariate matrices. The challenge we faced at this stage was therefore not so much how to set up the model as it was how to get the sampler to converge to the posterior distribution. The approaches we identified as effective were simplifying the model (for example, by setting some elements in the latent correlation matrix to be zero, or by excluding outcome-covariates) and including instrumental variables in the submodels. While the former is most effective, there may be a cost, since this approach may involve substantive changes to the postulated causal structure. If the analyst does decide to simplify the model, and there isn't interest in the effect of a particular outcome-covariate, excluding it from some submodels is likely to be more palatable than severing the latent link. This is because conditional independence of response variables given the latent variables (assumed if we exclude an outcome-covariate) is probably more reasonable than the assumption that the outcome-covariate is exogenous (assumed if we set the latent correlation to be zero). As an example, in Journal Article 6 (Chapter 11) we accommodated the relationship between number of embryos and fragmentation of individual embryos using only the correlated latent variables between the corresponding submodels. This component of the model was therefore similar to the joint models of cluster size and subunit-specific responses described by Dunson, et al. (2003). Instrumental variables (or, at least, variables that we assume have a substantive effect in the submodel in which they appear, and sufficiently negligible effects elsewhere to warrant exclusion) also facilitated the estimation of the model parameters. Due to the multilevel nature of IVF, where interventions are targeted both at women and at their oocytes and embryos, a number of plausible instrumental variables are available. Whether eggs are injected or mixed with sperm or vitro cannot possibly affect the patient directly, for example. Weakness or even outright invalidity of the instruments remains a possibility. The former arises where there is low correlation with the response in the submodel, while the latter arises if the instrument itself is endogenous (that is, the so-called 'back door criterion' is not satisfied). If injecting eggs with sperm is only minimally associated with embryo quality, then the instrument is weak. If clinicians decide to inject on the basis of prognostic patient characteristics, the instrument is invalid. Neither of these scenarios are fatal in the endogenous response model, since the correlated latent variables accommodate unmeasured confounding. Even when instrumental variables are included however, convergence of the models can be slow, necessitating moderately (although not excessively) long runtimes of several days to obtain large effective sample sizes. However, in Journal Article 5 (Chapter 10), we used real data to show that the endogenous response model can lead to substantively different conclusions compared to outcome regression, demonstrating the practical benefit that comes with the additional effort.

We further demonstrated the utility of the endogenous-response method by using it to investigate the effect of ovarian stimulation on uterine receptivity, distinct from indirect effects by way of oocytes and embryos. This suggested that embryo implantation was less likely when a patient received higher doses of gonadotropins, as was the likelihood of an embryo being carried to term. This result coincided with findings from previous, methodologically limited studies addressing the same topic and confirmed what was predicted on the basis of current understanding. However, it would be useful to confirm that the effect can be replicated in other datasets. One possibility would be to fit a version of the model in the dataset we used in Journal Article 4 (Chapter 9), where we quantified the sources of variation in ovarian response to stimulation. This dataset has additional information about ovarian reserve measures, which are used to determine dose. These variables may also be independently predictive of cycle outcome. In Journal Article 6 (Chapter 11), where these variables were not available, we relied upon our latent variable structure to soak up any confounding relating to ovarian reserve. The Journal Article 4 dataset does not have embryo-level outcome data however, meaning that we would not be able to consider dose effects on individual embryos. This is not a crucial feature for the purpose of answering the question of whether or not COS affects transfer outcomes, although an understanding of how dose affects embryo quality elucidates the complex relationship between gonadotropins and the final outcome. If convergent results were obtained from both datasets, this would strengthen the evidence for the effect. In addition, the paired analyses would together constitute a useful case study of the robustness of the endogenous-response method.

This latter point touches on a glaring limitation of the present work. Given the time requirement to fit these models, it has not been possible to evaluate their properties systematically through simulation studies. If a model takes eg: 2 days to fit, 1000 iterations would take five and a half years. Simulation studies have therefore not been possible, beyond some minimum testing of our models by way of fitting to small numbers of simulated datasets. In this regard, we largely stand upon the shoulders of those who have investigated the properties of these approaches mathematically (eg: Heckman, 1976, Heckman, 1978, Heckman, 1979, Terza, 1998 , Terza, 2009) and by simulation (eg: McCulloch, 2008). These studies have invariably focussed on simpler examples however, for example containing just two response variables. With the computational resources available, we would have been restricted to simulations based on similarly simplistic (compared to our actual examples) scenarios, which would lack relevance. More generally, a recent review of methodology to deal with unmeasured confounding suggested that there is relatively little work comparing the performance of different methods (Streeter et al., 2017).

Accordingly, we would urge that the methods we present are not implemented mindlessly, but are rather integrated into a process of thoughtful model building (controlling for confounding with measured variables as far as possible), model checking and sensitivity analysis. Of course, this should always be the case with data analysis. An example where mindless application of the method without sufficient understanding of the data generating process could potentially lead to error comes from Journal Article 6 (Chapter 11). The model in the article is set up to estimate the effect of varying dose on uterine receptivity, and we advise against interpretation of the other model covariates. For example, parameters relating to effects of embryo quality and double embryo transfer on embryo implantation are negative. A naïve interpretation would take this to indicate that higher quality embryos, and embryos transferred in duplicate, are less likely to implant. However, blastocyst stage embryos (those which have been cultured for five days) are usually transferred in singleton, and these are more likely to implant compared to cleavage stage embryos (which have been cultured for three). Further adjustment for day of transfer (blastocyst or cleavage stage) removes the apparent disadvantage of higher quality embryos and of double transfer on implantation. Because the models contain many parameters, a carefree approach to interpretation could result in many spurious conclusions.

We implemented the models in the Bayesian software RStan (Stan Development Team, 2014) primarily on pragmatic grounds. The software is flexible, and removes the need to write custom sampling algorithms. In addition, the Bayesian approach allows for stable estimation of parameters in complex models, since priors can be used to direct the sampler away from implausible values. In principle, this sort of prior regularisation can protect against overfitting a model to data (McElreath, 2015). However, it is important that the exclusion of implausible values is not prioritised over the inclusion of plausible ones. In Journal Article 5 (Chapter 10), we did not get this balance quite right. Our prior relating to the intercept term in our double embryo transfer submodel was too strong, underfitting the data (McElreath, 2015). This became apparent when we checked the model using draws from the posterior predictive distribution; the model slightly but systematically underestimated the sample mean, corresponding to the proportion of cycles where two embryos were transferred. In retrospect, it is quite obvious that a different prior should, in general, be appropriate for intercepts compared to other terms in the model, since these terms are on a different scale. This also highlights the need for robust model checking. Accordingly, in Journal Article 6 (Chapter 11), where we created and fitted a model in this framework to investigate the effects of ovarian stimulation on the uterine environment, we established that the estimated treatment effect was robust to informative prior specification for the model intercepts.

While we have provided Stan code with the models, the implementation of these methods will remain out of reach for most of those conducting IVF data analysis. One solution would be to produce an R package to simplify the process for the user. However, while our methods leverage what is, broadly speaking, the fixed, sequential structure of IVF, the particular submodels to include in any particular implementation will vary according to the research question. Creating a package capable of handling such a wide variety of models with minimal demands on the user might not be feasible. For now, researchers should modify the examples and code provided to produce bespoke models for their particular problems. We have shown that models of this complexity can be accommodated within Stan. We anticipate, but have not confirmed, that it will be possible to produce non-Bayesian analogues of many of the models we present using the Stata package gllamm (Generalised Linear Latent and Mixed Models), although whether or not it will be possible to fit the models without using prior regularisation is currently unclear. Shared parameter or ‘factor analysis’-style formulations of joint models (section 7.4.4), rather than formulations based on correlated latent variables, might fit more naturally within the GLLAMM framework. Michael Crowther’s upcoming meganreg (Multivariate Extended Generalised Linear and Non-linear Mixed Effects Models) package will accommodate a wide array of complex joint models, but will not accommodate responses defined at different levels of a hierarchy (personal communication). Accordingly, these methods will remain off-limits for all but the most competent analysts in the immediate future. Given our discussion outlining the risks of using complex methods without sufficient understanding, some would argue that the barrier to entry is a boon.

As we arrive at the conclusion of the thesis, there are many avenues for future research. Questions remain about the implications of model misspecification for these methods, particularly with respect to the posited latent variable distribution. This is by definition unobserved, and so cannot be checked against data. The posited distribution has implications for the assumed missing data mechanism. In our endogenous response models, we draw a complete vector of latent variables (with one element corresponding to each submodel) for each cycle regardless of when it was terminated. The latent variable distribution is the same for all cycles, both complete and incomplete, so that the patterns of confounding are assumed to be the same. It might be beneficial to relax the model assumptions by allowing the latent variable distribution to vary according to the stage of dropout. For example, cycles ending prior to transfer could have a distinct (and smaller) correlation matrix underlying their (smaller) set of submodels, compared to those who underwent all stages of treatment. The latent variables could then be fitted as mixture distributions (eg: Komarek, et al., 2010). This would bear some resemblance to the pattern mixture models of Little (1993). There may however be difficulties in implementation which we have not anticipated in the present discussion.

Our modelling work has primarily focussed on developing or extending methods for answering mechanistic or explanatory questions about the fresh IVF cycle. We have not yet considered the possibility of extending the models to cover frozen transfers (where stored, frozen embryos are thawed and transferred, usually after a failed fresh attempt).

Although we are yet to consider the matter in detail, we anticipate that it will be possible to extend the models we have described by adding further conditional submodels relating to implantation and live birth outcome after the frozen transfer procedure. We have also not extended the models to account for the clustering of repeated cycles within patients. By introducing patient-level random effects, a three-level structure (embryos within cycles within patients) could be modelled (Goldstein, 2003).

Another area for future research is the value of multistage modelling techniques for prediction problems. We gave an example in the discussion of Journal Article 5 (Chapter 10), where we used the correlated latent variable model to predict the probability that a patient would have a safe ovarian response and go on to have a baby. It is unclear whether outcome regression or endogenous response models would offer any advantage for problems such as these. We note at present that existing prediction methods offer black box prediction of the overall outcome of treatment given baseline characteristics. The ability to predict both the clinical outcome of the cycle and the route by which the patient arrives might assist in clinical decision making. These comments remain speculative for now.

We have not attempted to reconcile these methods with formal systems of causal inference. We have casually referred to ‘effects’ of non-manipulable variables, and have not given formal definitions of direct and indirect effects. For some problems, it might be useful to reformulate the models in a causal mediation framework. For example, in Journal Article 6 (Chapter 11), we have concentrated on estimating the direct effects of gonadotropin dose on responses throughout the model. A mediation analysis would allow us to quantify how much of the overall effect of COS is mediated through the quantity and quality of embryos produced. Pearl (2011) has described a mediation formula for nonlinear models, which can be fitted as a series of regression models. The details are yet to be considered fully.

Throughout, we have attempted to ensure that the methods we have developed are clinically relevant and have attempted to answer clinically important research questions. Our analyses in Journal Articles 4 and 6 (Chapters 9 and 11) are anticipated to have clinical value. The first suggests that the scope for tailoring the initial dose of gonadotropin to the individual on the basis of known markers is likely to be limited, due to the modest amount of explicable variation in ovarian response. Given the general enthusiasm of the IVF industry for new products which can be marketed to patients, together with the current fervour for personalised medicine, this sobering finding represents an important caveat. A Cochrane review of trials of personalised ovarian stimulation using ovarian reserve markers is underway (Lensen, et al., 2017).

Journal Article 6 (Chapter 11) suggests that clinicians should be mindful that the likelihood of embryo implantation and of term gestation might decrease with cumulative exposure to gonadotropin during COS. Since implantation failure is the most common reason for treatment failure, strategies to mitigate the adverse effects of COS have the potential to greatly improve success rates. Elective frozen embryo transfers (eFET, where all embryos are frozen for later transfer) and the transfer of blastocysts are two treatment strategies which might reduce the consequences of COS by giving the uterine environment time to recover. These strategies have disadvantages however. Both potentially reduce the pool the embryos available for transfer, since some embryos will perish before reaching blastocyst stage (Glujovsky, et al., 2016) and some will not survive the thawing process (Maheshwari and Bhattacharya, 2013). Cochrane reviews on these topics indicate that there is currently insufficient evidence to determine the effectiveness of eFET (Wong, et al., 2017),while suggesting an advantage of a blastocyst transfer policy as compared to a cleavage stage policy (Glujovsky, et al., 2016). One reason for this might be that waiting until day 5 for blastocysts to develop weeds out inferior embryos, which perish before this time. Acquiring greater understanding of these complex issues is a current goal of IVF research, and we anticipate that the methods we have outlined will be useful for this purpose.

The methods we have developed in this thesis are bespoke to IVF. We have identified treatment stages which appear in most IVF cycles, and have developed appropriate submodels for each of these. Any particular research question might require augmentation to this base set of submodels. At the time of writing, we have not identified other complex, multistage treatments which might be investigated using these approaches.

12.3 Closing Remarks

Maity and colleagues referred to the stages of the IVF cycle as 'failure opportunities' (Maity, et al., 2014). They were referring to the possibility of outright treatment failure, but this thesis demonstrates that they could just as well have been talking about the potential for methodological blunders. Because it is a multistage treatment, there are many options for reporting and analysis of IVF data, and this means that there are many opportunities to get things wrong. We would urge consumers and producers of IVF statistics to think particularly carefully about denominators, since these determine which patients are excluded from consideration. Where exclusions are not made clear, IVF success rates have the potential to be misleading. Producers of IVF statistics have a responsibility to mitigate these concerns by using intuitive, transparent outcome measures. There are additional problems with denominator-related exclusions in RCTs, since non-randomised comparisons do not guarantee valid inferences. The risk of misinterpretation is heightened when a non-randomised comparison is made in an RCT compared to an observational study design, since little appreciation of or control for confounding usually appears in the former. And of course, the perception of many researchers is that because the data arose from an RCT, any analysis of it must be trustworthy. RCTs are, after all, the gold standard in research. While we are sympathetic to the idea that different numerators are appropriate for different research questions, the variety currently appearing in the literature is excessive. Precise prespecification in a study protocol is essential, and peer reviewers should be wary of flexible outcome definitions.

The multistage treatment structure also results in a need for tools to address mechanistic research questions. These are important in the design of complex interventions, but are fiendishly difficult to answer correctly. The multiplicity of causal pathways encompassing eggs, embryos, selection and the uterine environment, as well as the complex patterns of confounding, means that standard methods are usually not sufficient for the task. We have developed methods for modelling multistage IVF data, and have provided code for the use of the applied statistician. Hopefully this makes the methods accessible enough for the competent analyst to use while keeping naïve users out of the loop. If members of the latter group had access to the models, the concern is that they might start dividing each of a multitude of parameters by its standard error, comparing each ratio against a t-distribution (with how many degrees of freedom?) and embarking on the fishiest of fishing expeditions. If our main contribution to the literature was to provide a new machine for the production of falsehoods, this would not be a good outcome. We hope instead that we have set the stage for conservative, high-quality structural analyses of the IVF cycle, and, on balance, created more opportunities for methodological success than for failure.

12.4 References for Chapter 12.

Allegra A, Marino A, Volpes A, Coffaro F, Scaglione P, Gullo S, La Marca A. A randomized controlled trial investigating the use of a predictive nomogram for the selection of the FSH starting dose in IVF/ICSI cycles. *Reprod Biomed Online* 2017.

Altman DG. The Scandal of Poor Medical-Research. *Bmj* 1994;308: 283-284.

Balen A, Regan L, Avery S, Braude P, Cooke I, Dugdale G, Seenan S, Khalaf Y, Rutherford A, Brinsden P *et al.* Re: Lack of evidence for interventions offered in UK fertility centres. *Bmj* 2017.

Bird SM, David C, Farewell VT, Harvey G, Tim H, Peter C. Performance indicators: good, bad, and ugly. *J R Stat Soc Ser A Stat Soc* 2005;168: 1-27.

Broer SL, Dolleman M, Opmeer BC, Fauser BC, Mol BW, Broekmans FJM. AMH and AFC as predictors of excessive response in controlled ovarian hyperstimulation: a meta-analysis. *Hum Reprod Update* 2011;17: 46-54.

Broer SL, Dolleman M, van Disseldorp J, Broeze KA, Opmeer BC, Bossuyt PMM, Eijkemans MJC, Mol BW, Broekmans FJM, Grp I-ES. Prediction of an excessive response in in vitro fertilization from patient characteristics and ovarian reserve tests and comparison in subgroups: an individual patient data meta-analysis. *Fertil Steril* 2013a;100: 420-+.

Broer SL, van Disseldorp J, Broeze KA, Dolleman M, Opmeer BC, Bossuyt P, Eijkemans MJC, Mol BWJ, Broekmans FJM, Grp IS. Added value of ovarian reserve testing on patient characteristics in the prediction of ovarian response and ongoing pregnancy: an individual patient data approach. *Hum Reprod Update* 2013b;19: 26-36.

Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* 2003;59: 521-530.

Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* 2013.

Glujovsky D, Farquhar C, Quinteiro Retamar AM, Alvarez Sedo CR, Blake D. Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. *Cochrane Db Syst Rev* 2016.

Goldstein H. 3-level Models and more Complex Hierarchical Structures *Multilevel Statistical Models, 4th Edition*. 2003, pp. 73-110.

Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. *Stat Model* 2009;9: 173-197.

Gueorguieva R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat Model* 2001;1: 177-193.

Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. *J Am Stat Assoc* 2001;96: 1102-1112.

Gueorguieva RV, Sanacora G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat Med* 2006;25: 1307-1322.

Heckman JJ. Common Structure of Statistical-Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Ann Econ Soc Meas* 1976;5: 475-492.

Heckman JJ. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* 1978;46: 931-959.

Heckman JJ. Sample Selection Bias as a Specification Error. *Econometrica* 1979;47: 153-161.

Higgins JP, Green S. *Cochrane handbook for systematic reviews of interventions*, 2011. John Wiley & Sons.

Khan K. The CROWN Initiative: Journal editors invite researchers to develop core outcomes in women's health. *Midwifery* 2014;30: 1147-1148.

Komarek A, Hansen BE, Kuiper EM, van Buuren HR, Lesaffre E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Stat Med* 2010;29: 3267-3283.

La Marca A, Grisendi V, Giulini S, Argento C, Tirelli A, Dondi G, Papaleo E, Volpe A. Individualization of the FSH starting dose in IVF/ICSI cycles using the antral follicle count. *J Ovarian Res* 2013;6.

La Marca A, Papaleo E, Grisendi V, Argento C, Giulini S, Volpe A. Development of a nomogram based on markers of ovarian reserve for the individualisation of the follicle-stimulating hormone starting dose in in vitro fertilisation cycles. *Bjog* 2012;119: 1171-1179.

Legro RS, Wu XK, Barnhart KT, Farquhar C, Fauser BCJM, Mol B, Conference HC, Comm S. Improving the Reporting of Clinical Trials of Infertility Treatments (IMPRINT): modifying the CONSORT statement. *Hum Reprod* 2014;29: 2075-2082.

Lensen SF, Wilkinson J, Mol BWJ, La Marca A, Torrance H, Broekmans FJ. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing IVF/ICSI. *Cochrane Db Syst Rev* 2017.

Little RJA. Pattern-Mixture Models for Multivariate Incomplete Data. *J Am Stat Assoc* 1993;88: 125-134.

Maheshwari A, Bhattacharya S. Elective frozen replacement cycles for all: ready for prime time? *Hum Reprod* 2013;28: 6-9.

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.

McCulloch C. Joint modelling of mixed outcome types using latent variables. *Stat Methods Med Res* 2008;17: 53-73.

McElreath R. Statistical Rethinking. Texts in Statistical Science. 2015. CRC Press.

Myatt L, Redman CW, Staff AC, Hansson S, Wilson ML, Laivuori H, Poston L, Roberts JM, COLAB GPC. Strategy for Standardization of Preeclampsia Research Study Design. *Hypertension* 2014;63: 1293-1301.

Nyboe Andersen A, Nelson SM, Fauser B, Garcia-Velasco JA, Klein BM, Arce JC. Individualized versus conventional ovarian stimulation for in vitro fertilization: A multicenter, randomized, controlled, assessor-blinded, phase 3 noninferiority trial. *Fertil Steril* 2017;107: 387-396.

Pearl J. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. 2011. CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.

Penman R, Heller G, Tyler J. Modelling IVF Data using an Extended Continuation Ratio Random Effects Model *Proceedings of the 22nd International Workshop on Statistical Modelling*. 2007, Barcelona.

Popovic-Todorovic B, Loft A, Bredkjaeer HE, Bangsboll S, Nielsen IK, Andersen AN. A prospective randomized clinical trial comparing an individual dose of recombinant FSH based on predictive factors versus a 'standard' dose of 150 IU/day in 'standard' patients undergoing IVF/ICSI treatment. *Hum Reprod* 2003;18: 2275-2282.

Repro Tech Truths. www.reprotechtruths.org. Last accessed 21st September 2017.

Rustamov O, Pemberton PW, Roberts SA, Smith A, Yates AP, Patchava SD, Nardo LG. The reproducibility of serum anti-Mullerian hormone in subfertile women: within and between patient variability. *Fertil Steril* 2011;95: 1185-1187.

Rustamov O, Smith A, Roberts SA, Yates AP, Fitzgerald C, Krishnan M, Nardo LG, Pemberton PW. Anti-Mullerian hormone: poor assay reproducibility in a large cohort of subjects suggests sample instability. *Hum Reprod* 2012;27: 3085-3091.

Rustamov O, Smith A, Roberts SA, Yates AP, Fitzgerald C, Krishnan M, Nardo LG, Pemberton PW. The Measurement of Anti-Mullerian Hormone: A Critical Appraisal. *J Clin Endocr Metab* 2014;99: 723-732.

Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci* 2011;22: 1359-1366.

Stan Development Team. RStan: the R interface to Stan, Version 2.5.0. 2014.

Terza JV. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *J Econom* 1998;84: 129-154.

Terza JV. Parametric Nonlinear Regression with Endogenous Switching. *Economet Rev* 2009;28: 555-580.

van't Hooft J, Khan K. P-hacking can be avoided with core outcome sets: preterm birth research is ready to take this leap. *BJOG* 2017;124: 1017-1017.

Williamson PR, Altman DG, Blazeby JM, Clarke M, Gargon E. The COMET (core outcome measures in effectiveness trials) initiative. *Trials* 2011;12: A70.

Wong KM, van Wely M, Mol F, Repping S, Mastenbroek S. Fresh versus frozen embryo transfers in assisted reproduction. *Cochrane Db Syst Rev* 2017.