Human Reproduction, Vol.33, No.5 pp. 784–792, 2018
Advanced Access publication on April 4, 2018 doi:10.1093/humrep/dey070

human reproduction

OPINION

A practical blueprint to systematically study life-long health consequences of novel medically assisted reproductive treatments

Callista L. Mulder[1,†], Joana B. Serrano[1,†], Lisa A.E. Catsburg[1], Tessa J. Roseboom[2,3], Sjoerd Repping[1], and Ans M.M. van Pelt[1,*]

[1]Center for Reproductive Medicine, Amsterdam Research Institute Reproduction and Development, Academic Medical Centre, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands [2]Department of Obstetrics and Gynaecology, Amsterdam Reproduction and Development Research Institute, Academic Medical Centre, Meibergdeef 9, 1105 AZ, Amsterdam, The Netherlands [3]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Public Health Research Institute, Academic Medical Centre, Meibergdeef 9, 1105 AZ, Amsterdam, The Netherlands

*Correspondence address. Center for Reproductive Medicine, Amsterdam Research Institute Reproduction and Development, Academic Medical Centre, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. E-mail: a.m.vanpelt@amc.uva.nl

Submitted on December 22, 2017; resubmitted on February 27, 2018

ABSTRACT: In medicine, safety and efficacy are the two pillars on which the implementation of novel treatments rest. To protect the patient from unnecessary or unsafe treatments, usually, a stringent path of (pre) clinical testing is followed before a treatment is introduced into routine patient care. However, in reproductive medicine several techniques have been clinically introduced without elaborate preclinical studies. Moreover, novel reproductive techniques may harbor safety risks not only for the patients undergoing treatment, but also for the offspring conceived through these techniques. If preclinical (animal) studies were performed, efficacy and functionality the upper hand. When a new medically assisted reproduction (MAR) treatment was proven effective (i.e. if it resulted in live birth) the treatment was often rapidly implemented in the clinic. For IVF, the first study on the long-term health of IVF children was published a decade after its clinical implementation. In more recent years, prospective follow-up studies have been conducted that provided the opportunity to study the health of large groups of children derived from different reproductive techniques. Although such studies have indicated differences between children conceived through MAR and children conceived naturally, results are often difficult to interpret due to the observational nature of these studies (and the associated risk of confounding factors, e.g. subfertility of the parents), differences in definitions of clinical outcome measures, lack of uniformity in assessment protocols and heterogeneity of the underlying reasons for fertility treatment. With more novel MARs waiting at the horizon, there is a need for a framework on how to assess safety of novel reproductive techniques in a preclinical (animal) setting before they are clinically implemented. In this article, we provide a blueprint for preclinical testing of safety and health of offspring generated by novel MARs using a mouse model involving an array of tests that comprise the entire lifespan. We urge scientists to perform the proposed extensive preclinical tests for novel reproductive techniques with the goal to acquire knowledge on efficacy and the possible health effects of to-be implemented reproductive techniques to safeguard quality of novel MARs.

Key words: health assessment / preclinical / MAR / offspring / transgenerational inheritance / DOHaD

Introduction

Since the birth of Louise Brown in 1978, many medically assisted reproduction (MAR) treatments have been introduced into clinical care, including IVF (Steptoe and Edwards, 1978), ICSI (Palermo *et al.*, 1992) and use of sperm collected through testicular sperm extraction (TESE) (Craft *et al.*, 1993; Devroey *et al.*, 1995). Even though at that time fundamental studies had been performed that suggested that

[†]These authors contributed equally to this article.

IVF may be effective to treat subfertility patients (Biggers, 2012), fertilization outside of the human body was initially a matter of intense debate among scientists, the media and the general public. People were concerned about the consequences of these treatments both on scientific, religious and moral grounds. One of the major concerns was the health of the individuals conceived *in vitro*. Renowned scientists feared for children born with severe malformations (Ramsey, 1972a, b; Marantz Henig, 2004) and urged for more safety studies prior to clinical application of IVF. But since Louise Brown was born healthy, concerns about the safety of IVF faded away quickly. Hitherto, acceptance of more refined MAR techniques, including ICSI and TESE in the following years occurred with less societal upheaval, and more importantly, without preclinical safety testing.

As defined by Harper *et al*. (2012), ‘every procedure involving application to the human body should be defined as experimental until adequate scientific evidence is provided regarding its safety and efficacy’. In our opinion, novel reproductive technologies should be no exception to this, especially since they might not only affect the patient, but the offspring and grand-offspring too (Barker, 2004; Daxinger and Whitelaw, 2012a; Van Otterdijk and Michels, 2016).

Despite the increasing use of MAR, knowledge about long-term effects of MAR in the parent, offspring or even grand-offspring is limited. Health consequences in MAR-derived offspring are primarily studied in retrospect, i.e. often many years after implementation of the treatment into clinical care. The first long-term health study on the health of IVF children was published 12 years after its clinical implementation (Morin *et al.*, 1989). Since the first study, many papers have been published describing effects of IVF on the offspring. Although children conceived through MAR are predominantly healthy (Hart and Norman, 2013a, b), there is also evidence of an increased risk of congenital abnormalities (Rimm *et al.*, 2004; Davies *et al.*, 2012; Wen *et al.*, 2012; Hansen *et al.*, 2013), low birthweight (Schieve *et al.*, 2002; Helmerhorst *et al.*, 2004; Jackson *et al.*, 2004; Dumoulin *et al.*, 2010; Kleijkers *et al.*, 2016), growth and developmental deviations (Koivurova *et al.*, 2003; Kai *et al.*, 2006; Ceelen *et al.*, 2009) and increased levels of cardiovascular and metabolic markers later in life (Ceelen *et al.*, 2008, 2009; Sakka *et al.*, 2010; Hart and Norman, 2013b). This could suggest that IVF children are at increased risk of developing cardiovascular and metabolic diseases. Due to their relevatively young age such effects have not been assessed in humans. However, animal models have not only shown that IVF significantly reduced lifespan in mice exposed to a high fat diet (Rexhaj *et al.*, 2013), but also that this procedure induces significant epigenetic modifications that can have developmental and metabolic effects on the offspring (Mahsoudi *et al.*, 2007; Chen *et al.*, 2014; Feuer and Rinaudo, 2017). Despite the important findings in these studies, interpretation of the overall effect is often complex due to the use of various methods to perform IVF, missing controls and various clinical definitions of outcome measures (e.g. birth defects). Moreover, data on the health effects of MAR-derived offspring are not collected uniformly, since standard protocols are lacking (Fauser *et al.*, 2014).

In the light of the present data on the health of IVF children it is remarkable that preclinical (animal) testing is not standard in the field of reproductive medicine. Without doubt, the aim of all fertility clinics is to provide a safe way for prospective parents to have a healthy child. However, we cannot neglect that MAR has become part of huge commercial market worldwide where most treatments are performed in the private sector, which is dependent on economic competition. Subfertile patients are willing to undergo invasive, burdensome treatments to have a biologically own child, even if the safety of these treatments is unknown. In many cases, if a treatment is not offered or is illegal in the patients’ home country, they travel across borders to obtain medical help in conceiving (McKelvey *et al.*, 2009; Shenfield *et al.*, 2010). Moreover, the desire for a child of their own sometimes exceeds the desire for a healthy child (Hendriks *et al.*, 2014). Ultimately, fertility clinics may perceive the pressure to deliver high success rates and to offer state-of-the art techniques, even when these may not have been validated properly. And since preclinical testing is not required for novel MAR, clinicians and scientists are not obliged to perform costly, but critical preclinical tests.

Currently, more elaborate MAR are being developed, including the creation of artificial gametes through induced pluripotent stem cells and the genetic manipulation of embryos. Although there is indeed broad consensus on the need to conduct preclinical safety studies, a framework on how this should be done for MARs is currently lacking. Hence, there is an urgent need for standardized methodologies that help researchers to perform health assessment of the MAR offspring besides proof-of-concept for therapeutic effect (Hyun *et al.*, 2008; Freedman and Inglese, 2014) before clinical implementation.

In our opinion, systematic safety testing prior to introducing a novel MAR technology into clinical care should be the standard. An ideal paradigm for hypothesis-driven research has been proposed earlier (Harper *et al.*, 2012), starting with fundamental research where the physiology and development of gametes and embryos is scrutinized, which in many cases precedes the initial idea of a novel technique. Once an idea of a novel technique has been spurred, a series of lifelong experiments involving animals, including small rodents and preferably also larger animals, should follow after which the technique may be tested in human gametes or embryos donated for research. If results are promising a small scale clinical study should precede a large multi-center clinical study. These studies are prospective in nature and include long-term follow-up of the children derived from this novel MAR. In the end clinical and cost-effectiveness should be assessed.

In this paper we provide a blueprint for systematically studying the health of MAR-derived offspring in a systematic and hypothesis-driven experimental design. We specifically propose a series of experiments in mouse that may help to predict the safety of novel MAR technologies after clinical application in humans.

The mouse as a practical model to assess health in MAR offspring

For now, it is still not possible to study the general health of MAR-derived offspring *in vitro* before clinical implementation, because it requires actual reproduction which is only feasible *in vivo* and an alternative is not available. To translate optimally the findings from preclinical safety and efficacy studies of new MARs to humans, animal with corresponding physiological resemblance would be preferable to minimize possible species-specific effects. After all, no animal model can fully recapitulate the anatomy and physiology of a human being. However, one can attempt to make the translation from animal to human as straightforward as possible.

Non-human primate models show the highest genetic similarity to man, with a genetic divergence of a mere 1–3% in chimpanzees, gorilla and orangutan (Chen and Li, 2001). Due to their high similarity with human (i.e. cognitive, biological and physiological developments), it is unethical and forbidden under EU legislation (Directive 2010/63/EU) to work with these animals for research purposes. Other non-human primates can be used for proof-of principle studies (Schlatt *et al.*, 1999; Hermann *et al.*, 2012). However, due to their relatively long lifespan and low availability, studies of long-term consequences of new MARs in other non-human primates as well as large domestic animals would require long study periods covering sometimes even decades (Boerjan *et al.*, 2000; Ceelen and Vermeiden, 2001; Vodička *et al.*, 2005; Zheng *et al.*, 2014; Lorenzen *et al.*, 2015).

Because of their relatively short lifespan and their high reproduction rates, rodents are the most practical animal model for transgenerational studies, specifically small rodents such as mice. The mouse shows a low genetic divergence of ~10% when compared to humans (Waterston *et al.*, 2002). The genetic similarity of inbred mice is of value for easier interpretation of experimental data, thereby decreasing variability in the results and reducing the number of animals required. Besides the scientific benefits, the mouse, as an animal model for man, enables the study of age-related diseases in a relatively short period of time, demands low maintenance costs and has balanced ethical arguments (Santulli *et al.*, 2015). Moreover, since the sexual maturity of the mouse is reached at the age of 6–8 weeks, it allows for a transgenerational study design in a relative short time span, which makes it less time consuming when assessing health of MAR-derived offspring. Therefore, we suggest the use of a mouse model as a first step in preclinical MAR research.

Important stages of life

In general, the average lifespan of a mouse is 18–24 months depending on the strain, while the mean life expectancy of a human being is currently 71 years (World Health Organization, 2016). It has been calculated that 1 human year would be comparable with 9 mouse days, however, one must consider the relative pace in which the organism develops (Dutta and Sengupta, 2016). It has been estimated that in the first month, a mouse matures 150 times faster than a human being, therefore a mouse of 28 days is comparable to a child of 11.5 years of age (The Jackson Laboratory, 2017). For this procedure, we have divided life in two stages, child development (day of birth up to 28 days of age) and adulthood (3–18 months of age) (Fig. 1). The transition from childhood to adulthood is marked as the ability to reproduce. Development and health assessment tests are performed in these two stages, including a histopathological health assessment at 18 months, which is comparable to the health status of a 56-year-old human being (The Jackson Laboratory, 2017). Even though 56 years of human age had been suggested just to be the onset of ageing, we would recommend not to prolong beyond this age of the test animals,

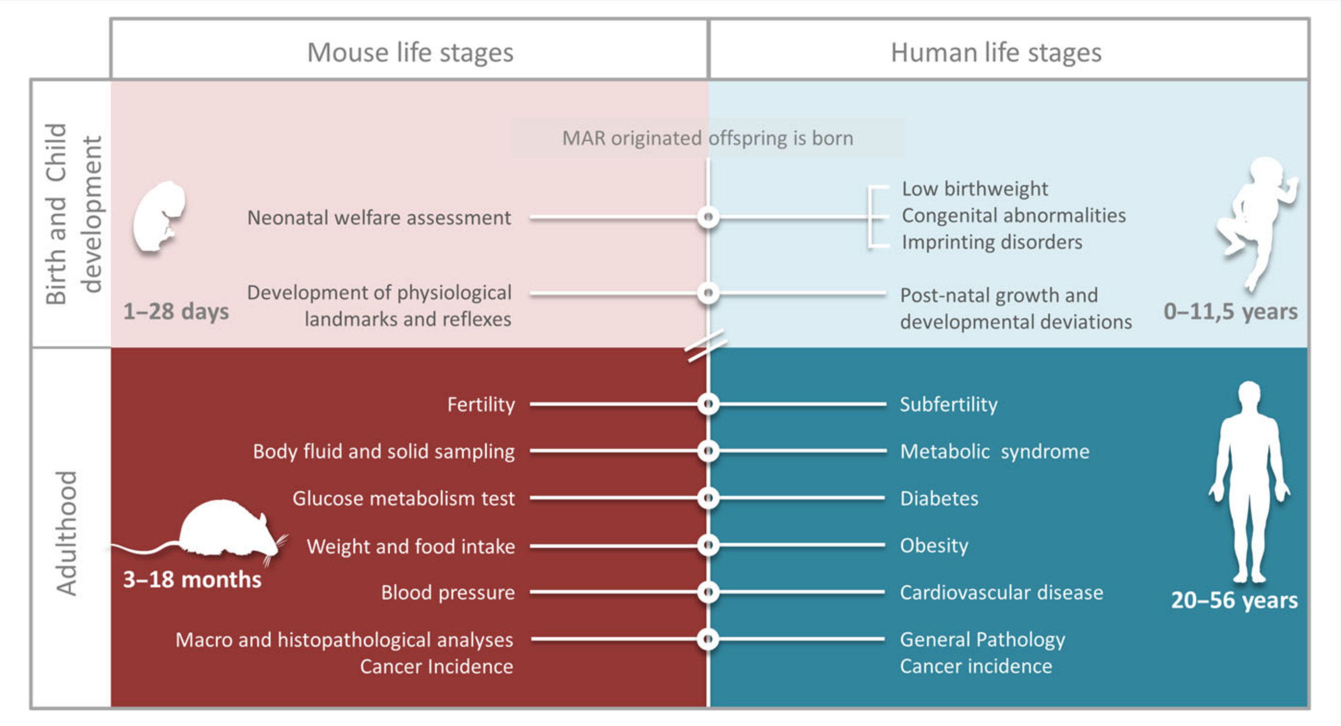

Figure 1 Physiological tests in a timeline from birth to death in mouse versus human lifespan. The two life stages are child development and adulthood. Each stage includes examples of physiological tests with higher relevance according to the age period. The life cycle of a mouse is depicted in red and human in blue and the corresponding ages between the two species are presented below the illustrations for each stage. The extensive study of different developmental stages in mouse allows for the investigation of age-related developmental and metabolic disorders in human. (Images are adapted from Servier Medical Art by Servier (http://smart.servier.com/) and modified by the authors under the following terms: CREATIVE COMMONS Attribution 3.0 Unported (CC BY 3.0).)

due to a rapid increase in age-related discomfort in the animals. By performing this series of tests we are able to assess health throughout the entire life-span, which will help us to predict the health status of future MAR-derived children.

Assessing childhood health and development in a mouse model

Litter characteristics

To investigate the effects of novel MAR procedures on the offspring, testing should be performed to identify physical and behavioral deviations in the neonates (van der Meer *et al.*, 2001; Turgeon and Meloche, 2009). Once pups are born, the litter should be analyzed immediately to establish general parameters including litter size, number of live births, birth weight and length and congenital anomalies. These could include grossly visible anomalies, including limb defects, alterations in body shape or position. Congenital anomalies may be immediately apparent at birth, including spina bifida, microcephaly and orofacial clefting. However, one must keep in mind that many other anomalies will not be visible at the day of birth and will present at a later stage of development. Typically, congenital heart defects present when oxygen levels lowered, resulting in cyanosis of the animal (van der Meer *et al.*, 2001; Hood, 2005; Turgeon and Meloche, 2009), therefore, later testing for congenital anomalies is advised.

When analyzing congenital anomalies it is important to take stillbirths and perinatal deaths into account. However, perinatal death is quite common in mouse breeding and it is known that the mother is prone to cannibalize the cadavers of the pups, thereby leaving the researcher unaware of the existence of these newborns. This makes it difficult to parse out whether pups are stillborn, die because of congenital disease or were subjected to infanticide (active killing by the mother) (Weber *et al.*, 2013). To reduce stress to the mother and the litter new gloves should be used while handling the animals (and switch between litters) while some pups are left with the dam when the tests are being performed, preferably out of the vicinity of the cage. In this light, documenting litter size is of paramount importance, since a decreased litter size may suggest an increased fetal resorption, stillbirths or congenital anomalies. This argues again for immediate analyses once the pups are born and we would advise to store deceased pups at 4°C for necropsy (Hood, 2005). Importantly, subjective litter effects can be observed, which will have a great effect in the results and account for variation in the data (Lazic and Essioux, 2013). Experimental design can be further improved taking into account statistical approaches that deal with litter-to-litter variations (Lazic and Essioux, 2013).

Physical development

A well-designed follow-up during the first 4 weeks of life for each individual newborn mouse is crucial to assess childhood development. During these first 28 days of life, multiple developmental milestones can be checked to investigate physical development and general behavior of the pups. A delay in acquiring pivotal physical landmarks during neonatal life suggests a delay in development (Hood, 2005). These physical landmarks include, amongst others, date of opening eyes and ears, fur growth and incisor eruption. Weight and length are monitored to allow comparison to a standard weight curve. In wild-type naturally conceived animals, the mouse is born naked with closed ears and eyes. Between Days 2 and 4 hair starts to appear, between Days 13 and 14 the ears open, and between Days 14 and 15 the eyes also open (van der Meer *et al.*, 1999, 2001). Around Day 16, when the eyes are completely functional, the pups will begin to eat solid food but nursing can continue for 1 more week. After 3 weeks of age the pups resemble an adult mouse except for their size and differentiation of the sexual organs (Silver, 1995).

When assessing development of MAR-derived offspring on the mentioned aspects, specific time periods for testing are suggested based on time intervals empirically determined from wildtype offspring. The pups originated from MARs techniques may develop these landmarks sooner or later than the control pups. Based on a pilot experiment, we suggest the following time intervals and frequency of these tests opening of eyes (Days 10–17); opening of ears (Days 10–17); hair

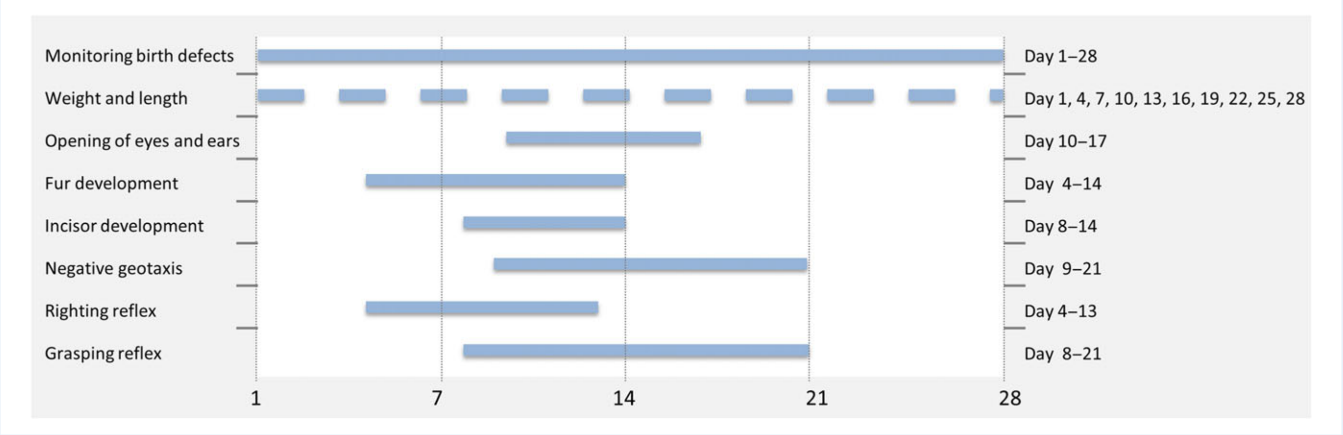

Figure 2 Planning of neonatal tests in developing mice. Schematic representation of multiple morphological evaluations and reflex-ontogeny tests that can be performed in mice between 1 and 28 days of age. The dashed line corresponds to intermittent weight and length measurements while the full lines indicates continued analyses.

growth (Days 4–14); tooth development (Days 8–13); and periodic weighting and length measurements (every 3 days during the first 28 days) (Fig. 2).

Functional development

Mice and humans are both altricial species with significant neonatal ontogenetic landmarks of the nervous system. Therefore neurodevelopmental disorders in human, such as growth restriction and delays in the appearance of developmental milestones, can be a modeled in the mouse (Hill *et al.*, 2008). We propose that therapeutic and translational safety of new MAR techniques in human infants are assessed with functional neurodevelopmental outcomes of the neonatal mice derived from MAR. The animals can be subjected to a number of reflex-tests that help to determine the reflex-ontogeny and partial behavioral development of the mice, e.g. righting reflex, grasp reflex, negative geotaxis (van der Meer *et al.*, 1999; Hood, 2005; Feather-Schussler and Ferguson, 2016). The tests begin before the onset of the developmental landmark in wildtype and continue daily until each animal in the litter meets the criterion (Fig. 2). For each day of testing, the results are presented in sequential levels evaluating their response in each test: 0 (behavior or response is absent), 1 (primitive response), 2 (a clear but not yet mature response) or 3 (a mature and full response in all aspects of execution such as coordination or strength) (van der Meer *et al.*, 1999). Ultimately, in a well-powered experiment (see Methodology issues) the means are compared to identify potential statistically significant differences between the tested MAR and naturally conceived animals.

Based on our experience and that of others, we propose the inclusion of the surface righting reflex from Days 4 to 13, which is the ability of regaining the normal position after the mouse pup is placed on its back (Feather-Schussler and Ferguson, 2016). The ideal response is when the animal rights itself immediately demonstrating labyrinthine reflexes and complex coordinated action involving muscles in the neck, trunk and limbs. (van der Meer *et al.*, 1999; Heyser, 2004; Hood, 2005; Hill *et al.*, 2008) and can be compared to the skills needed for a human infant to roll over. Motor coordination and labyrinthine reflexes can also be assessed with the negative geotaxis reflex test in young pups from Days 9 to 21 (Fig. 2). Mice are placed in a slope facing downwards and a delayed or failed response to turn upwards could indicate deficits in coordination, balance or vestibular input (van der Meer *et al.*, 1999, 2001; Hill *et al.*, 2008; Ferguson and Bailey, 2013). Between Days 8 and 21 (Fig. 2) we suggest the grasping test that evaluates fine motor skills of the mice (van der Meer *et al.*, 1999). This reflex appears in humans at birth and disappears around 5–6 months of age. Grasping deficits indicate impairment of the nervous system (Feather-Schussler and Ferguson, 2016) and therefore are essential to take into account.

Assessing health during adulthood in a mouse model

Fertility

During adulthood, fertility via natural mating should be tested when MAR-derived animals (F1) reach the breeding age to create an F2 generation. Depending on the studied MAR, a third generation (F3) should be created in order to find true transgenerational effects (Daxinger and Whitelaw, 2012b). This is because in the case of exposures of gestating mothers (F0), the fetus (F1) and its embryonic developing germline (F2) are also exposed. Therefore, a third generation is required to find the phenotypic effects of transgenerational persistent inheritance (Van Otterdijk and Michels, 2016). However, exposures to male or female gametes (F0) (before gestation) directly affect the F1 generation, so a second generation is sufficient to assess transgenerational persistent effects. Health assessment of the F2 and F3 generations again involves assessment of health and development of childhood and adulthood (Fig. 1). Selection of breeding animals from each litter (F1) should be at random when no significant differences are observed in physical traits between the litter mates (OECD, 2001). If there are systematic differences between the littermates we suggest the use of stratified randomization of the animals (see Methodology) to avoid allocation bias that could influence the outcome in the different treatment groups (Kao *et al.*, 2008).

Behavior and learning

Behavioral testing may be a valuable asset to acquire knowledge on stress, depression, learning and memory. For example spatial learning can be assessed using the Water Morris maze in mice (D'Hooge and De Deyn, 2001; Barnhart *et al.*, 2015). However, since it is known that the performance of the mice in behavioral tests is dependent on their strain, careful experimental design and local expertise is imperative (Upchurch and Wehner, 1988; Moy *et al.*, 2007).

Cardiovascular and metabolic assessment

As the mice reach adulthood (Fig. 1), metabolic analyses as well as cardiovascular risk factors are obviously more significant. There is evidence that the IVF procedure induces increased risk of hypertension and diabetes in mice (Watkins *et al.*, 2007; Rexhaj *et al.*, 2013) and there is a growing body of evidence in humans suggesting that blood pressure is increased, glucose tolerance is reduced and insulin resistance increased (Ceelen *et al.*, 2008, 2009; Sakka *et al.*, 2010; Hart and Norman, 2013b). Therefore, during adulthood, metabolic tests should be performed to assess the effect of MAR technologies on the originated offspring. These tests can among others include investigation of body fluids (blood, saliva, urine, feces), blood pressure analysis, glucose tolerance test and insulin resistance, weight and food intake, body composition and physical activity (Fig. 1). Since non-invasive scanning procedures, like MRI, are becoming available for animal science as well, one could opt for scanning of various organs including brain at various ages.

General macro and histopathological analyses

Ultimately, in this mouse model, parents and offspring should be anatomically examined post-mortem, and bodily fluids and organs collected for pathological analyses and compared to natural conceived animals. By performing a comprehensive necropsy, one is able to identify diseases that are not directly apparent in the animal, including malignant or benign growths. We would advise to perform this necropsy at a fixed age, as it allows for direct comparison between MAR-derived animals and controls, without having to correct for age. The necropsy can be viewed as a final evaluation of health. Analyses can include life span, cancer incidence and general macro- and histopathological examination (Mulder *et al.*, 2018).

Methodology

As in any experiment or (clinical) trial, proper experimental design is the key to a successful study. Especially in preclinical translational animal studies it is of paramount importance to create a solid foundation to one's experiment and mimic the intended clinical therapy as closely as possible taking into account the development in time and potential clinical relevant artefact of the model and the method used with correct controls.

Power calculation

In general, we recommend a pilot study to gain more knowledge on effect sizes. In some cases, it may be acceptable to perform elaborate literature review to acquire this information, although we do recommend a pilot study if practically feasible. Knowledge gained from a pilot animal study helps to perform a proper power calculation to determine the sample size, which will not only facilitate the ability to draw conclusions, but is (or should be) often required for ethical approval of animal experiments. This process will help to assess the feasibility of the reproductive technique of choice, optimize the technique and to gain more knowledge on important statistical parameters. Of course one has to also determine the primary outcome of the study in order to be able to perform any power calculation. We would advise to always consult a biostatistician when to choose the most suitable primary outcome (Festing and Altman, 2002). Stratification methods should be applied when factors such as gender, weight and age are expected to affect these outcomes (Indrayan and Holt, 2016).

Control group

From our experience, many factors influence both the feasibility and outcome of a study. Firstly, the choice of proper control groups is essential to a well-designed experiment. A control breeding line, originating from breeding couples with identical strain, age, diet and housing conditions, allows for accurate comparison of the MAR with control natural conception. Depending on the research question other control groups are sensible. For instance, when assessing an adaptation to an IVF protocol we would recommend to always include a standard IVF control group, or if a new technique requires IVF the control group should also be derived via IVF and not natural conception. Practically, we urge other researchers to perform all tests in parallel and within a similar timeframe as for the experimental group. We do not recommend to rely only on literature based reference values, as important testing parameters might be influenced by subtle variations in environmental conditions such as housing or food composition (Smith *et al.*, 2016).

Double blind assessment

Randomization and blinding are of key importance to provide unbiased results. The experiments should be designed to enable blinded analyses to minimize selection and observer bias during the experiments and upon outcome assessment. Additionally, random allocation of the animals to an experimental group and/or randomization of the outcome assessment reduces confounding (Macleod *et al.*, 2015). Another general concern that should be considered includes reporting the characteristics of the study, which are described here in Methodology, preferably following the ARRIVE guidelines (Kilkenny *et al.*, 2010).

Mouse strain

The choice of mouse strain can greatly influence the feasibility of a model. Since Dr Clarence Cook Little began inbreeding of wildtype mice in 1909, a vast array of mouse strains have been developed, C57BL/6 being the most widely used in science and the first to have its genome sequenced (Waterston *et al.*, 2002). However, many genetic variants exist between inbred strains. A comparative analysis of the genomes of 17 widely used inbred strains revealed many functional variants (Keane *et al.*, 2011). Moreover, fertility associated differences exists between inbred mouse strains (Braden, 1959; Shire and Bartke, 1972). The strain of mice used can depend on the animal model that is best suited for the specific MAR being tested. Phenotypic research on mouse models should be performed during design of the experimental to account for the differences between strains (https://phenome.jax.org/). Therefore, it is of importance to choose the mouse strain with care by performing a literature search and preferably a pilot study prior to initiating the project.

Breeding

Depending on the MAR technique that is being tested, breeding can be done in timed mating or via continuous mating. Through timed mating one can calculate the time of gestation and monitor the fetal growth during pregnancy. Timed pregnancy can involve embryo transfer or over-night breeding while the female is in oestrus. In the case of embryo transfer it is of importance to transfer an equal number of embryos to the left and right uterine horns. Precise recording of this will help to identify reduced litter size. When opting for continuous mating, where the male and female are always together, timed pregnancy is not feasible. In the latter case, it is important that the male is separated from the female before she gives birth. This will prevent a back-to-back pregnancy, which is suggested to be associated with decreased health of both litters because of undernutrition (van Zutphen *et al.*, 2001). For identification and proper follow-up per individual animal, newborns can be tattooed on the day of birth by an intradermal injection of India ink in the palms of the paws (Hood, 2005). At 15 days of age it is allowed to permanently identify the animals through an earcut.

Optimized reproductive performance

The procedures should be designed to optimize reproductive performance to obtain the required number of offspring and to reduce variability in neonatal testing. More specifically, environmental factors can greatly affect the robustness of specific parts of health assessment. Experimental refinement may involve having a 14 h light period and the inclusion of reversed day/night cycle in order to perform tests during the dark period when the mice are active (Roedel *et al.*, 2006; Pritchett and Taft, 2007). The set of experiments that we propose explores behavioral responses and morphological development of the mice, therefore stress and discomfort may introduce variability in the experimental results (Lupien *et al.*, 2009). In addition, housing can influence the behavior of mice, as the use of individually ventilated cages (IVC) has been shown to have behavioral effects in C57BL/6J mice, such as anxiety-like effects (Logge *et al.*, 2013). Moreover, it was demonstrated that housing in IVC racks slightly reduces the number of pups per female in DBA/2 mice when compared to conventional open racks (Tsai *et al.*, 2003) which is a crucial parameter in assessing adverse effect of MARs.

Experimental planning

Ultimately, we urge researchers to design the experiments thoroughly, adhering to the 3R's (replacement, reduction, refinement) principles of animal welfare. For the time being, the use of animals is our sole option when studying the health of MAR-derived offspring, since offspring cannot be produced without actual reproduction. However, by performing pilot studies and having a thorough design of the experimental settings (correct animal model, accounting for environmental conditions, blinding, randomization, correct use of control groups, etc.) one can reduce the number of animals required while decreasing variation in the results. The National Center for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) developed a helpful free online tool to design preclinical studies: the Experimental Design Assistant (EDA) (https://eda.nc3rs.org.uk/) (Percie du Sert *et al.*, 2017).

Translation of animal health assessment to clinical care

The ultimate and primary goal of this set of experiments is to predict the health status of offspring derived from novel MAR. We urge researchers to publish their findings in peer-reviewed open access journals, whether they point out health risks or not. And if health risks are found, it remains to be decided if these risks are acceptable in order to proceed to clinical trials.

In our opinion, preclinical animal safety testing should be a prerequisite for obtaining ethical approval for a Phase I clinical trial. The use of well-designed animal models is of great importance for the reproductive field as animal research will give us the information we need to proceed with confidence towards clinical trials. The verdict whether a MAR can be introduced should be reached by an (inter)national ethical committee, as these decisions require a balanced opinion of, amongst others, scientists, policy makers and ethicists (Hendriks *et al.*, 2018).

Preclinical animal research in general is used to identify health risks of a novel treatment, however, true health effects in human remain to be identified in human beings. Therefore, prospective follow-up of the treated parent and children from a novel MAR is key to verify safety in human.

Of course, besides assuring safety of novel therapies, this type of research has a high scientific value as well. During this set of experiments a vast amount of tissues and cell types are harvested, which can aid to unravel underlying mechanisms. Additionally, this blueprint also allows to study if exposures of the parents germline have prejudicial effects on the unexposed offspring (F2 or F3, depending on the therapy) by transgenerational inheritance (Daxinger and Whitelaw, 2012a; Van Otterdijk and Michels, 2016). For this purpose, we recommend that tissue or cell types from the three germ layers are acquired. After death, sperm can be isolated from the epididymis and follicles collected from the ovary to study as a proxy for the next generation.

Concluding remarks

In this article, we present a blueprint to encourage researchers to pursue preclinical testing of safety of newly developed MAR technologies. Using an animal model such as the mouse allows for a lifelong study of the health of MAR-derived offspring in a relative short period, in which multiple stages of development are included. Periodic tests in each lifestage can be adapted for a variety of outcomes and the tissues of interest can be used for histopathologic examinations and multiple omics analyses. Moreover, the mouse model enables the efficient study of multiple generations which is required to investigate the potential of inheritance of transgenerational effects of newly developed MARs, before going for clinical implementation. Given the contradictory literature on the effects of MAR currently used in the clinic in retrospect, the preclinical assessment of novel MAR techniques on inbred mice without a pathological reproductive background is in our opinion imperative.

Authors' roles

A.M.M.P, S.R., C.L.M. and J.B.S. initiated and designed the outline of this opinion paper. C.L.M., L.A.E.C., J.B.S., A.M.M.P. and T.J.R. were involved in designing and optimizing the proposed experimental protocol. C.L.M. and J.B.S. drafted the original article. All authors critically reviewed and revised the article and approved the final version.

Funding

Stichting Kinderen Kankervrij (KiKa86) and ZonMw (TAS116003002).

Conflict of interest

The authors report no financial or other conflict of interest relevant to the subject of this article.

References

Barker DJ. The developmental origins of adult disease. *J Am Coll Nutr [Internet]* 2004;**23**:588S–595S.

Barnhart CD, Yang D, Lein PJ. Using the Morris water maze to assess spatial learning and memory in weanling mice. *PLoS One* 2015;**10**:1–16.

Biggers J. IVF and embryo transfer: historical origin and development. *Reprod Biomed Online* 2012;**25**:118–127.

Boerjan ML, Daas JHG Den, Dieleman SJ. Embryonic origins of health: long term effects of IVF in human and livestock. *Theriogenology* 2000;**53**: 537–547.

Braden AWH. Strain differences in the morphology of the gametes of the mouse. *Aust J Biol Sci* 1959;**12**:65–71.

Ceelen M, Van Weissenbruch MM, Prein J, Smit JJ, Vermeiden JPW, Spreeuwenberg M, Van Leeuwen FE, Delemarre-Van De Waal HA. Growth during infancy and early childhood in relation to blood pressure and body fat measures at age 8–18 years of IVF children and spontaneously conceived controls born to subfertile parents. *Hum Reprod* 2009; **24**:2788–2795.

Ceelen M, van Weissenbruch MM, Vermeiden JP, van Leeuwen FE, Delemarre-van de Waal HA. Cardiometabolic differences in children born after in vitro fertilization: follow-up study. *J Clin Endocrinol Metab [Internet]* 2008;**93**:1682–1688.

Ceelen M, Vermeiden JP. Health of human and livestock conceived by assisted reproduction. *Twin Res* 2001;**4**:412–416.

Chen FC, Li WH. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* 2001;**68**:444–456.

Chen M, Wu L, Zhao J, Wu F, Davies MJ, Wittert GA, Norman RJ, Robker RL, Heilbronn LK. Altered glucose metabolism in mouse and humans conceived by IVF. *Diabetes* 2014;**63**:3189–3198.

Craft I, Bennett V, Nicholson N. Fertilising ability of testicular spermatozoa. *Lancet* 1993;**342**:864.

Davies MJ, Moore VM, Willson KJ, Van Essen P, Priest K, Scott H, Haan EA, Chan A. Reproductive technologies and the risk of birth defects. *N Engl J Med* 2012;**366**:1803–1813.

Daxinger L, Whitelaw E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet [Internet]* 2012a;**13**: 153–162.

Daxinger L, Whitelaw E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Publ Gr [Internet]* 2012b;**13**:153–162. Nature Publishing Group.

Devroey P, Liu J, Nagy Z, Goossens A, Tournaye H, Camus M, Van Steirteghem A, Silber S. Pregnancies after testicular sperm extraction and intracytoplasmic sperm injection in non-obstructive azoospermia. *Hum Reprod [Internet]* 1995;**10**:1457–1460.

Dumoulin JC, Land JA, Van Montfoort AP, Nelissen EC, Coonen E, Derhaag JG, Schreurs IL, Dunselman GA, Kester AD, Geraedts JP *et al*. Effect of in vitro culture of human embryos on birthweight of newborns. *Hum Reprod [Internet]* 2010;**25**:605–612.

Dutta S, Sengupta P. Men and mice: relating their ages. *Life Sci* 2016;**152**: 244–248. Elsevier Inc.

D'Hooge R, De Deyn PP. Applications of the Morris water maze in the study of learning and memory. *Brain Res Rev* 2001;**36**:60–90.

Fauser BC, Devroey P, Diedrich K, Balaban B, Bonduelle M, Delemarre-van de Waal HA, Estella C, Ezcurra D, Geraedts JPM, Howles CM *et al*. Health outcomes of children born after IVF/ICSI: a review of current expert opinion and literature. *Reprod Biomed Online* 2014;**28**:162–182. Reproductive Healthcare Ltd.

Feather-Schussler DN, Ferguson TS. A battery of motor tests in a neonatal mouse model of cerebral palsy. *J Vis Exp* 2016;**117**:e53569. doi: 10.3791/53569.

Ferguson DR, Bailey MM. Reproductive performance of mice in disposable and standard individually ventilated cages. *J Am Assoc Lab Anim Sci* 2013; **52**:228–232.

Festing MFW, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J* 2002;**43**:244–258.

Feuer SK, Rinaudo PF. Physiological, metabolic and transcriptional post-natal phenotypes of in vitro fertilization (IVF) in the mouse. *J Dev Orig Health Dis* 2017;**8**:403–410.

Freedman LP, Inglese J. The increasing urgency for standards in basic biologic research. *Cancer Res* 2014;**74**:4024–4029.

Hansen M, Kurinczuk JJ, Milne E, Klerk N, de, Bower C. Assisted reproductive technology and birth defects: a systematic review and meta-analysis. *Hum Reprod Update [Internet]* 2013;**19**:330–353.

Harper J, Cristina Magli M, Lundin K, Barratt CLR, Brison D. When and how should new technology be introduced into the IVF laboratory? *Hum Reprod* 2012;**27**:303–313.

Hart R, Norman RJ. The longer-term health outcomes for children born as a result of ivf treatment. Part II—mental health and development outcomes. *Hum Reprod Update* 2013a;**19**:244–250.

Hart R, Norman RJ. The longer-term health outcomes for children born as a result of ivf treatment: Part I—general health outcomes. *Hum Reprod Update* 2013b;**19**:232–243.

Helmerhorst FM, Perquin DA, Donker D, Keirse MJ. Perinatal outcome of singletons and twins after assisted conception: a systematic review of controlled studies. *Br Med J [Internet]* 2004;**328**:261.

Hendriks S, Dancet EA, Meissner A, Van der Veen F, Mochtar MH, Repping S. Perspectives of infertile men on future stem cell treatments for nonobstructive azoospermia. *Reprod Biomed Online [Internet]* 2014; **28**:650–657.

Hendriks S, Vliegenthart R, Repping S, Dancet EAF. Broad support for regulating the clinical implementation of future reproductive techniques. *Hum Reprod* 2018;**33**:39–46.

Hermann BP, Sukhwani M, Winkler F, Pascarella JN, Peters KA, Sheng Y, Valli H, Rodriguez M, Ezzelarab M, Dargo G *et al*. Spermatogonial stem cell transplantation into rhesus testes regenerates spermatogenesis producing functional sperm. *Cell Stem Cell [Internet]* 2012;**11**:715–726. Elsevier Inc.

Heyser CJ. Assessment of developmental milestones in rodents. *Curr Protoc Neurosci/Editor Board, Jacqueline N Crawley [et al.]* 2004; **Chapter 8**:Unit 8.18.

Hill JM, Lim MA, Stone MM. Developmental milestones in the newborn mouse. *Neuropept Tech* 2008;**39**:131–149.

Hood RD (ed). *Developmental and Reproductive Toxicology: A Practical Approach*. Boca Raton, FL: CRC Press, 2005.

Hyun I, Lindvall O, Ährlund-Richter L, Cattaneo E, Cavazzana-Calvo M, Cossu G, De Luca M, Fox IJ, Gerstle C, Goldstein RA *et al*. New ISSCR guidelines underscore major principles for responsible translational stem cell research. *Cell Stem Cell* 2008;**3**:607–609.

Indrayan A, Holt MP. *Concise Encyclopedia of Biostatistics for Medical Professionals*. Boca Raton, FL: Chapman and Hall/CRC, 2016.

Jackson RA, Gibson KA, Wu YW, Croughan MS. Perinatal outcomes in singletons following in vitro fertilization: a meta-analysis. *Obs Gynecol* 2004;**103**:551–563.

Kai CM, Main KM, Andersen AN, Loft A, Chellakooty M, Skakkebaek NE, Juul A. Serum insulin-like growth factor-I (IGF-I) and growth in children born after assisted reproduction. *J Clin Endocrinol Metab* 2006;**91**:4352–4360.

Kao LS, Tyson JE, Blakely ML, Lally KP. Clinical research methodology I: introduction to randomized trials. *J Am Coll Surg* 2008;**206**:361–369.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M *et al*. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 2011;**477**:289–294.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE Guidelines for Reporting Animal Research. *PLoS Biol* 2010;**8**:e1000412.

Kleijkers SHM, Mantikou E, Slappendel E, Consten D, Van Echten-Arends J, Wetzels AM, Van Wely M, Smits LJM, Van Montfoort APA, Repping S *et al*. Influence of embryo culture medium (G5 and HTF) on pregnancy and perinatal outcome after IVF: a multicenter RCT. *Hum Reprod* 2016; **31**:2219–2230.

Koivurova S, Hartikainen AL, Sovio U, Gissler M, Hemminki E, Järvelin MR. Growth, psychomotor development and morbidity up to 3 years of age in children born after IVF. *Hum Reprod* 2003;**18**:2328–2336.

Lazic SE, Essioux L. Improving basic and translational science by accounting for litter-to-litter variation in animal models. *BMC Neurosci* 2013;**14**:37.

Logge W, Kingham J, Karl T. Behavioural consequences of IVC cages on male and female C57BL/6J mice. *Neuroscience* 2013;**237**:285–293.

Lorenzen E, Follmann F, Jungersen G, Agerholm JS. A review of the human vs. porcine female genital tract and associated immune system in the perspective of using minipigs as a model of human genital Chlamydia infection. *Vet Res* 2015;**46**:1–13. Veterinary Research.

Lupien SJ, McEwen BS, Gunnar MR, Heim C. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nat Rev Neurosci* 2009;**10**:434–445.

Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, Hirst T, Hemblade R, Bahor Z, Nunes-Fonseca C *et al*. Risk of bias in reports of in vivo research: a focus for improvement. *PLoS Biol* 2015;**13**:1–12.

Mahsoudi B, Li A, O'Neill C. Assessment of the long-term and transgenerational consequences of perturbing preimplantation embryo development in mice 1. *Biol Reprod* 2007;**77**:889–896.

Marantz Henig R. *Pandora's Baby—How the First Test Tube Babies Sparked the Reproductive Revolution*. New York, USA: Cold Spring Harbor, 2004.

McKelvey A, David AL, Shenfield F, Jauniaux ER. The impact of cross-border reproductive care or 'fertility tourism' on NHS maternity services. *BJOG An Int J Obstet Gynaecol* 2009;**116**:1520–1523.

Morin NC, Wirth FH, Johnson DH, Frank LM, Presburg HJ, Van de Water VL, Chee EM, Mills JL. Congenital malformations and psychosocial development in children conceived by in vitro fertilization. *J Pediatr* 1989;**115**:222–227.

Moy SS, Nadler JJ, Young NB, Perez A, Holloway LP, Barbaro RP, Barbaro JR, Wilson LM, Threadgill DW, Lauder JM *et al*. Mouse behavioral tasks relevant to autism: phenotypes of 10 inbred strains. *Behav Brain Res* 2007;**176**:4–20.

Mulder CL, Catsburg LAE, Zheng Y, Winter-Korver CMD, Daalen SKMV, Van Wely M, Pals S, Repping S, Pelt AMMV. Long-term health in recipients of transplanted in vitro propagated spermatogonial stem cells. *Hum Reprod* 2018;**33**:81–90.

OECD. *OECD Guidelines for the Testing of Chemicals, Section 4*. Paris, France: OECD Publishing, 2001.

Palermo G, Joris H, Devroey P, Van Steirteghem AC. Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. *Lancet [Internet]* 1992;**340**:17–18.

Percie du Sert NP, Bamsey I, Bate ST, Berdoy M, Clark RA, Cuthill IC, Fry D, Karp NA, Macleod M, Moon L *et al*. The experimental design assistant. *Nat Methods* 2017;**15**:1–2. Nature Publishing Group.

Pritchett KR, Taft RA. Reproductive Biology of the Laboratory Mouse. *Mouse Biomed Res* 2007, pp. 91–121. Elsevier.

Ramsey P. Shall we 'Reproduce'? I. The medical ethics of in vitro fertilization. *J Am Med Assoc* 1972a;**220**:1346–1350.

Ramsey P. Shall we 'Reproduce' II. Rejoinders and future forecast. *J Am Med Assoc* 1972b;**220**:1480–1485.

Rexhaj E, Paoloni-Giacobino A, Rimoldi SF, Fuster DG, Anderegg M, Somm E, Bouillet E, Allemann Y, Sartori C, Scherrer U. Mice generated by in vitro fertilization exhibit vascular dysfunction and shortened life span. *J Clin Invest* 2013;**123**:5052–5060.

Rimm AA, Katayama AC, Diaz M, Katayama KP. A meta-analysis of controlled studies comparing major malformation rates in IVF and ICSI infants with naturally conceived children. *J Assist Reprod Genet* 2004;**21**: 437–443.

Roedel A, Storch C, Holsboer F, Ohl F. Effects of light or dark phase testing on behavioural and cognitive performance in DBA mice. *Lab Anim* 2006;**40**:371–381.

Sakka SD, Loutradis D, Kanaka-Gantenbein C, Margeli A, Papastamataki M, Papassotiriou I, Chrousos GP. Absence of insulin resistance and low-grade inflammation despite early metabolic syndrome manifestations in children born after in vitro fertilization. *Fertil Steril* 2010;**94**:1693–1699.

Santulli G, Borras C, Bousquet J, Calzà L, Cano A, Illario M, Franceschi C, Liotta G, Maggio M, Molloy WD *et al*. Models for preclinical studies in aging-related disorders: one is not for all. *Transl Med @ UniSa* 2015;**13**: 4–12.

Schieve LA, Meikle SF, Ferre C, Peterson HB, Jeng G, Wilcox LS. Low. Low and very low birth weight in infants conceived with use of assisted reproductive technology . *N Engl J Med* 2002;**346**:731–737.

Schlatt S, Rosiepen G, Weinbauer GF, Rolf C, Brook PF, Nieschlag E. Germ cell transfer into rat, bovine, monkey and human testes. *Hum Reprod [Internet]* 1999;**14**:144–150.

Shenfield F, Mouzon J, de, Pennings G, Ferraretti AP, Nyboe Andersen A, Wert G, de, Goossens V. Cross border reproductive care in six European countries. *Hum Reprod* 2010;**25**:1361–1368.

Shire JG, Bartke A. Strain differences in testicular weight and spermatogenesis with special reference to C57BL-10J and DBA-2J mice. *J Endocrinol* 1972;**55**:163–171.

Silver LM. *Mouse Genetics: Concepts and Applications*. Oxford University Press, 1995.

Smith MM, Clarke E, Little CB. Considerations for the Design and Execution of Protocols for Animal Research and Treatment to improve reproducibility and standardization: 'DEPART well-prepared and ARRIVE safely.'. *Osteoarthr Cartil* 2016;**25**:1063–4584. Elsevier Ltd.

Steptoe PC, Edwards RG. Birth after the reimplantation of a human embryo. *Lancet [Internet]* 1978;**2**:366.

The Jackson Laboratory. Life Span as a Biomarker. 2017. Retrieved from https://www.jax.org/research-and-faculty/research-labs/the-harrison-lab/gerontology/life-span-as-a-biomarker#.

Tsai P-P, Oppermann D, Stelzer HD, Mähler M, Hackbarth H. The effects of different rack systems on the breeding performance of DBA/2 mice. *Lab Anim* 2003;**37**:44–53.

Turgeon B, Meloche S. Interpreting neonatal lethal phenotypes in mouse mutants: insights into gene function and human diseases. *Physiol Rev* 2009;**89**:1–26.

Upchurch M, Wehner JM. Differences between inbred strains of mice in Morris water maze performance. *Behav Genet* 1988;**18**:55–68.

van der Meer M, Baumans V, Hofhuis FM, Olivier B, van Zutphen BF. Consequences of gene targeting procedures for behavioural responses and morphological development of newborn mice. *Transgenic Res* 2001; **10**:399–408.

van der Meer M, Costa P, Baumans V, Olivier B, Van Zutphen B. Welfare assessment of transgenic animals: behavioural responses and morphological development of newborn mice. *Altern Lab Anim* 1999;**27**:857–868.

Van Otterdijk SD, Michels KB. Transgenerational epigenetic inheritance in mammals: how good is the evidence? *FASEB J* 2016;**30**:2457–2465.

van Zutphen LF, Baumans V, Beynen AC. Zutphen LF, van, Baumans V, Beynen AC (eds). *Principles of Laboratory Animal Science*, revised edn. Elsevier, 2001. ISBN 9780444506122.

Vodička P, Smetana K, Dvořánková B, Emerick T, Xu YZ, Ourednik J, Ourednik V, Motlík J. The miniature pig as an animal model in biomedical research. *Ann N Y Acad Sci* 2005;**1049**:161–171.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;**420**:520–562.

Watkins AJ, Platt D, Papenbrock T, Wilkins A, Eckert JJ, Kwong WY, Osmond C, Hanson M, Fleming TP. Mouse embryo culture induces changes in postnatal phenotype including raised systolic blood pressure. *Proc Natl Acad Sci* 2007;**104**:5449–5454.

Weber EM, Algers B, Hultgren J, Olsson IAS. Pup mortality in laboratory mice—infanticide or not? *Acta Vet Scand* 2013;**55**:83.

Wen J, Jiang J, Ding C, Dai J, Liu Y, Xia Y, Liu J, Hu Z. Birth defects in children conceived by in vitro fertilization and intracytoplasmic sperm injection: a meta-analysis. *Fertil Steril* 2012;**97**:1331–1337.e4. Elsevier Inc.

World Health Organization. *World Health Statistics 2016: Monitoring Health for the SDGs*. WHO Press, 2016.

Zheng Y, Zhang Y, Qu R, He Y, Tian X, Zeng W. Spermatogonial stem cells from domestic animals: progress and prospects. *Reproduction [Internet]* 2014;**147**:65–74.