THE APPLICATION OF MORPHOKINETIC ALGORITHMS TO PREDICT PLOIDY STATUS DURING ASSISTED CONCEPTION

By

THOMAS JAMES BAMFORD [redacted]

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Institute of Metabolism and Systems Research

College of Medical and Dental Sciences

University of Birmingham

-

November 2023

ABSTRACT

Aim: To undertake a series of studies to answer 5 key questions within assisted conception:

1. Are morphokinetic variables and morphological features associated with the ploidy status of pre-implantation human embryos?
2. Are artificial intelligence or machine learning algorithms superior to logistic regression for predicting ploidy status?
3. Are morphokinetic model risk scores associated with live birth and miscarriage?
4. Should clinical factors be incorporated into embryo selection models?
5. Are morphokinetic models better at prioritising a euploid embryo for transfer over morphological selection by a senior embryologist?

Methods: The above questions were addressed throughout four studies, first, a systematic review and meta-analysis investigated the association of ploidy status and abnormal cleavage, morphokinetic variables, fragmentation, multinucleation and embryo contraction. Second, a model development study collected data on the prognostic variables investigated in the systematic review from nine IVF clinics. Here, a sample of 8148 biopsied blastocysts was used to develop and compare 12 machine learning models to predict ploidy status. This was using four different algorithms, logistic regression, random forest classifier, extreme gradient boosting and deep learning. One model for each algorithm was built with euploidy as target outcome, a second with aneuploidy and a third using a smaller dataset which incorporated embryo Gardner's classification. Third, the best performing model was retrospectively externally tested on a total of 3587 single embryo transfers. This determined association between three different model derived aneuploidy risk scores (low, moderate and high) and live birth and miscarriage. The final study used a separate

cohort of 1958 biopsied blastocysts to compare the ability of morphokinetic models to rank euploid embryos first, given that these models will not be asked to classify embryos but only prioritise.

Results: Meta-analysis demonstrated that ten morphokinetic variables were significantly delayed in aneuploid embryos. It is uncertain whether the morphological components investigated have prognostic potential. On comparing 12 different models, logistic regression performed the best (AUC=0.61). Including predictors such as age resulted in no variability in the ranking within a patient's cohort of embryos. Incorporating morphological Gardner's classification resulted in no improvement in the discriminatory ability of the model. A 'morphokinetics only' approach was therefore investigated by adjusted logistic regression analysis that demonstrated the model was not associated with miscarriage when comparing the 'high' to the 'moderate risk' (OR 0.87; 95% CI 0.63-1.63; p=0.39) or 'high' to 'low risk' embryos (OR 1.07; 95% CI 0.79-1.46, p<0.63). However, an embryo deemed 'low risk' was significantly more likely to result in a live birth than those embryos graded 'high risk' (OR 1.95; 95% CI 1.65-2.25; p<0.001). The final cohort study reported that arbitrary embryo selection would rank a euploid embryo first 37% of the time, embryologist selection 39%, and the ploidy morphokinetic model 47% of the time.

Conclusions: Morphokinetic variables and the risk scores derived from morphokinetic models are significantly associated with ploidy status. Including predictors such as age, results in a clinically ineffective model; a 'morphokinetics only' approach is therefore advised. Logistic regression was the best performing algorithm in this dataset for predicting ploidy status, with aneuploidy as the target variable. The application of this model resulted in an improved chance of a euploid embryo being selected for transfer over that by a senior embryologist.

DEDICATION

First, I would like to dedicate this thesis to my husband, Cristiano, who has supported me without fail on all my ventures, academic or personal. Cristiano, you have supported every decision I have made throughout my career thus far without question, no matter how ambitious or challenging. Thank you for being my rock and closest friend. During this PhD you agreed to marry me, we became dog parents and we both took on demanding jobs. I will never forget how you stepped in to support me and our family while I worked many late nights and weekends on the projects that follow.

Secondly, to my parents, Enid and Kevin Bamford. This PhD wouldn't exist if it wasn't for you and your constant support for me. You have helped me cultivate a rewarding and fulfilling career. We have both had to navigate difficult paths and challenges for the first time, yet you never faltered. Thank you for the love and kindness you have shown.

ACKNOWLEDGEMENTS

Thank you to Professor Arri Coomarasamy for his supervision, wisdom, and guidance. No hour is too late or proposition too silly. It has been a real privilege to be supervised by someone so talented and intelligent. You have helped spark my interest in academia and helped me develop a passion for Reproductive Medicine.

I would also like to thank the clinicians at Care Fertility who both helped to acquire the funding for this PhD but also support me through it while training me in IVF. Those would include Professor Charles Kingsland, Dr Victoria Sephton, Dr Julija Gorodeckaja, and Dr Philip Lowe. This thank you should also be extended to my team at Care Manchester who were all so welcoming and have since been so supportive as I've grown into a new role.

Further to this I would also like to thank the senior embryologists at Care Fertility who introduced me to the topic of morphokinetics, supported the review and revisions of manuscripts and helped facilitate data collection. This would include Professor Alison Campbell, Dr Sue Montgomery, Dr Amy Barrie, and Rachel Smith. Thank you to Dr

Amy Barrie for being my second reviewer in what was a large systematic review and meta-analysis. I would also like to acknowledge the embryologists, Selina Young, Amy Evans, Megan Lockwood who helped with data collection and the ranking of embryos. This PhD would have been even more challenging and difficult if it wasn't for Rachel Smith, who helped me navigate and understand the complex databases within Care Fertility and how I could best collect, combine and validate this information.

Christina Easter and the statistics team at the University of Birmingham were instrumental in helping me construct a robust methodology for the creation and validation of the morphokinetic models. This is also extended to Professor Jean-Baptiste (director of the centre of computational biology) and Dr Justina Zurauskiene (lecturer in computational biology). Both of whom were pivotal in helping me understand the coding and application of machine learning algorithms. The University of Birmingham supercomputer team, 'BlueBEAR', were also instrumental in my understanding how to navigate python using Linux. In particular thank you to Aslam Ghumra in the IT department for his help and ability to fix any computational issues I had.

Importantly, thank you to the people who formed my projects institutional review board for study ethics approval. This included the chair, Mr Ektoras Georgiou, Mrs Oonagh Pickering and Ms Janine O'Rourke.

Finally, thank you to the Tommy's team for welcoming me into this team, often virtually, and in particular to my second supervisor, Dr Rima Dhillon-Smith. Rima has helped

support me through a difficult career decision and offered academic advice when reviewing work submitted for publication.

All of the work presented in this thesis is my own. Where contributions have been made by others, this has been detailed in the title page of that corresponding chapter.

ABBREVIATIONS AND DEFINITIONS

aCGH: Microarray-based comparative genomic hybridization

AC: Abnormal cleavage- When the embryo doesn't display normal cytokinesis or division patterns one would normally expect

AFC: Antral follicle count

AI: Artificial Intelligence

AMA: Advanced maternal age

AMH: Anti-mullerian hormone

ANN: Artificial neural networks

ART: Assisted Reproductive Technology

AUC: Area under the curve

BMI: Body mass index

Cc2: Time of the second cell cycle (t_3-t_4), from 2 to 3 cells

Cc3: Time of the third cell cycle (t_5-t_3), from 2 to 3 cells

CI: Confidence Interval

CITL: Calibration in the large

D3: day 3 biopsy

DL: Deep learning

DUC1: abnormal division after syngamy resulting in 3-4 blastomeres.

DUC2: abnormal cleavage at the 2-cell stage resulting in 5 or 6 blastomeres.

DUC3: abnormal cleavage at the 4-cell stage resulting in 9 blastomeres or more.

E2: Oestradiol level

FISH: Fluorescent in situ hybridisation

FSH: Follicle stimulating hormone

GnRH: Gonatrophin-releasing Hormone

GRADE: Grading of recommendations, assessment, development and evaluation guidelines

HCG: human chorionic gonadotropin

HFEA: Human fertilisation and embryology authority

ICC: Intraclass coefficient

ICM: Inner cell mass

ICSI: Intra-cytoplasmic sperm injection

IECV: Internal external cross validation

IQR: Interquartile range

IVF: In-vitro fertilisation

LB: Live birth model

LH: Luteinising Hormone

LR: Logistic Regression

M1: Fully compacted morula

M2: Morula where partial exclusion of cells observed

MK: Morphokinetic

MN: Multinucleation

NDCG: Normalised discounted cumulative gain

NGS: Next generation sequencing

O/E: Ratio of observed to expected values

OR: Odds ratio

PA: Previous child affected by aneuploidy

PGT-A: Pre-implantation genetic testing for aneuploidy

PGS: Pre-implantation genetic screening

PREFER: Predicting euploid for embryos in reproductive medicine (ploidy prediction model using morphokinetics and clinical variables)

PREFER-MK: Ploidy prediction model using only morphokinetic predictors

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis

QUIPS: Quality in Prognosis studies tool

RFC: Random forest classifier

RIF: Recurrent implantation failure

RM: Recurrent miscarriage

RC: Reverse cleavage which results from blastomere fusion.

RCT: Randomised controlled trial

RR: Relative risk

S2: Time of synchrony of second cell cycle (t_4-t_3) from 2 to 4 cells.

S3: Time of synchrony of third cell cycle (t_8-t_5) from 4 to 8 cells.

SD: Standard deviation

SNP: Single nucleotide polymorphisms

SSR: Surgical sperm retrieval

STROBE: Strengthening the reporting of observational studies in epidemiology

TESE: Testicular sperm extraction

t*n*: time from insemination to completed division of n cells.

tPB2: time to appearance of second polar body.

tPNf: Time when both pronuclei have faded.

tSC: Time from insemination to the start of compaction (when the first cells of the embryo join).

TLS: time lapse system

tM: Time from insemination to the formation of the morula, where all cells have undergone compaction and cell boundaries are unclear.

tSB: Time from insemination to the start of blastulation when the first signs of cavitation are visible.

tB: Time from insemination to the formation of a full blastocyst, when the blastocoele filled the embryo with <10% in diameter.

tEB: Time from insemination to expanded blastocyst; when the blastocyst had increased in diameter by more than 30% and the zona pellucida starts to thin

tHB: Time from insemination to hatching blastocyst, trophectoderm herniation through the zona pellucida is observed.

TVOR: transvaginal oocyte retrieval

USA: United states of America

UK: United Kingdom

XGBoost: Extreme gradient boosting

TABLE OF CONTENTS

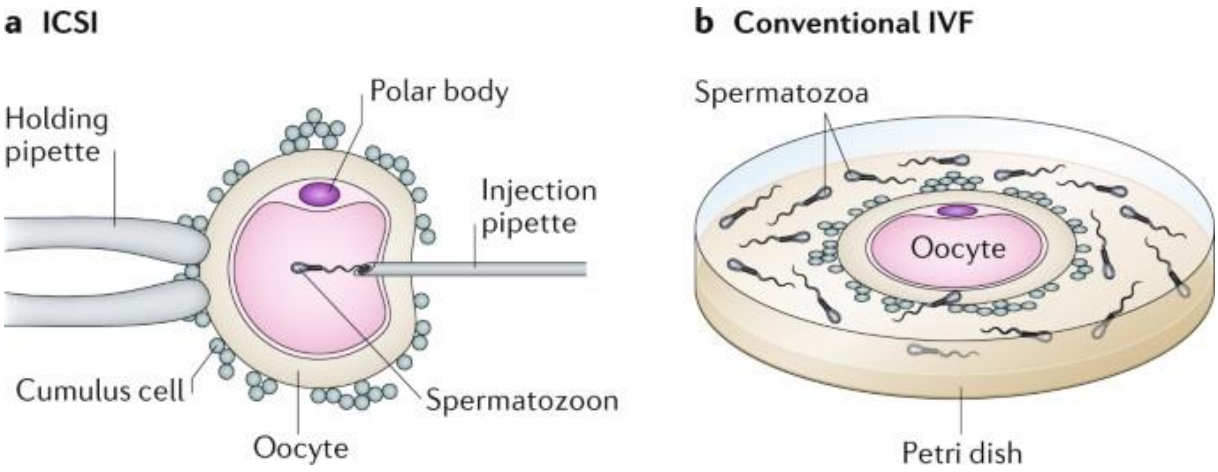
CHAPTER 1: INTRODUCTION

BACKGROUND

Assisted conception or assisted reproductive technology (ART) involves the stimulation of a patient's ovaries with gonadotrophins to induce multi-follicular growth. Follicles 13mm or more in size typically contain a mature oocyte; stimulation is therefore tailored to maximise the number of mature eggs retrieved by monitoring follicular growth with transvaginal ultrasound scans (Shapiro et al. 2022). Once a patient is deemed ready for an egg collection (traditionally when three follicles are 17mm or more), a 'trigger' injection of Human Chorionic Gonadotropin or Gonadotrophin-releasing hormone (GnRH) agonist (GnRH) is used to induce a surge of Luteinising Hormone (LH). This leads to oocyte maturation in-vivo whereby the oocyte completes meiosis I and stops at metaphase II until fertilisation, when meiosis II is complete (Seibel et al. 1982). In keeping with the timing of the LH surge, a transvaginal oocyte retrieval (TVOR) is performed 36 hours post trigger injection. This involves the use of a transvaginal pelvic ultrasound scan; a needle is passed through the vaginal wall and into each ovary to aspirate the follicular fluid. This fluid is then examined for oocytes which are then identified by the laboratory team. The eggs are either inseminated with prepared sperm or injected via ICSI (intracytoplasmic sperm injection) depending on semen analysis results (Figure 1). The fertilised eggs are then cultured and where possible, transferred into the uterus at the blastocyst stage on day five. There is a drive for single embryo transfers to avoid the complications associated with multiple pregnancy, therefore the best quality embryo is prioritised. The patient would commence progesterone luteal support prior to embryo transfer, normally undertaken five days following the TVOR. Any good quality spare embryos are

cryopreserved for future use in a subsequent cycle involving endometrial preparation with oestrogen and progesterone.

Figure 1: Insemination methods (used with permission from Esteves et al. 2018)

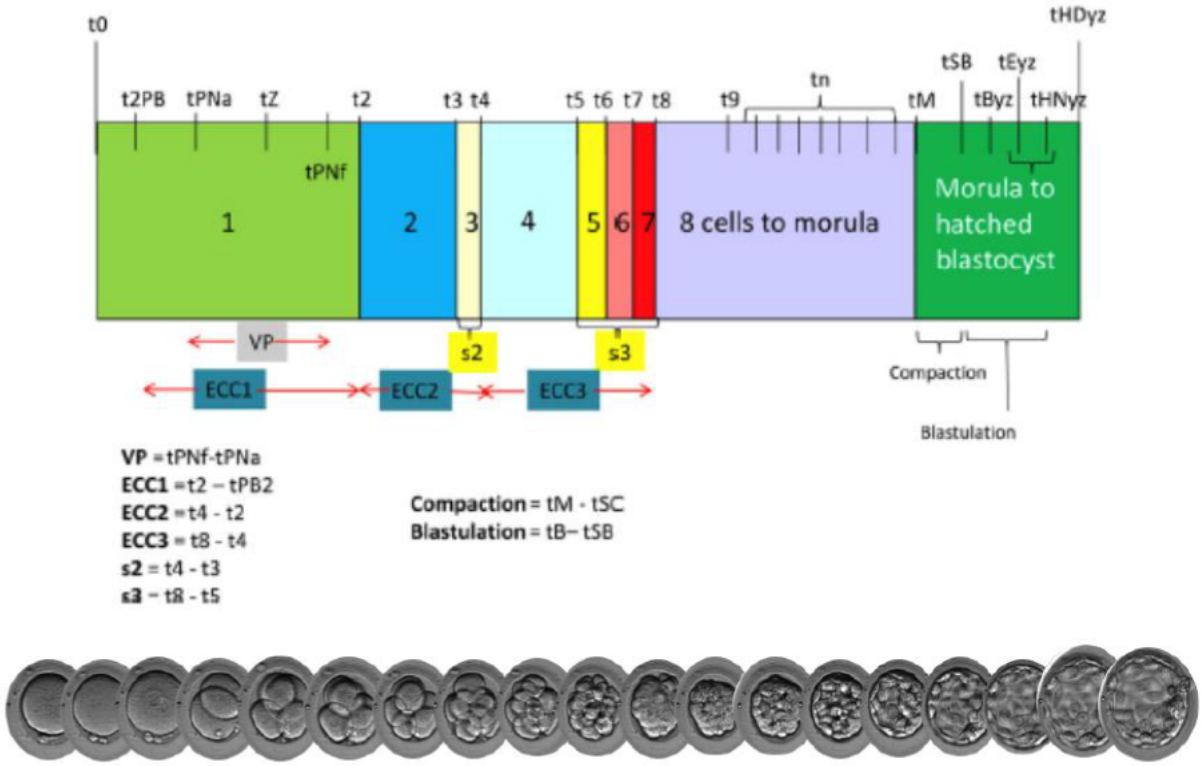


Many of the processes that dictate the success of the embryo transfer happen within the laboratory and ensuring the best embryo is selected. Historically, embryos are cultured and assessed using morphological evaluation at specific time points of development, examining the quality of the inner cell mass and trophectoderm (Cummins et al. 1986; Gardner et al. 2000; Alpha Scientists in Reproductive and Embryology 2011). This remains a subjective assessment despite consensus criteria; phenomena are also potentially missed between check points. The alternative of monitoring using a time lapse system (TLS) was first introduced in 1997 whereby frequent images are taken of early embryo development. This was extended to the blastocyst stage ten years later (Mio and Maeda 2008). These images are then annotated such that morphokinetic variables are created, for instance, time to four cells (t4) or start of compaction (tSC) (see list of definitions and Figure 2). Importantly, these observations are recorded without removing the embryo from the incubator, avoiding

changes in light, humidity, temperature, pH, and gases. This data can then be used to create algorithms for embryo selection to predict blastocyst formation, implantation and most recently, ploidy status.

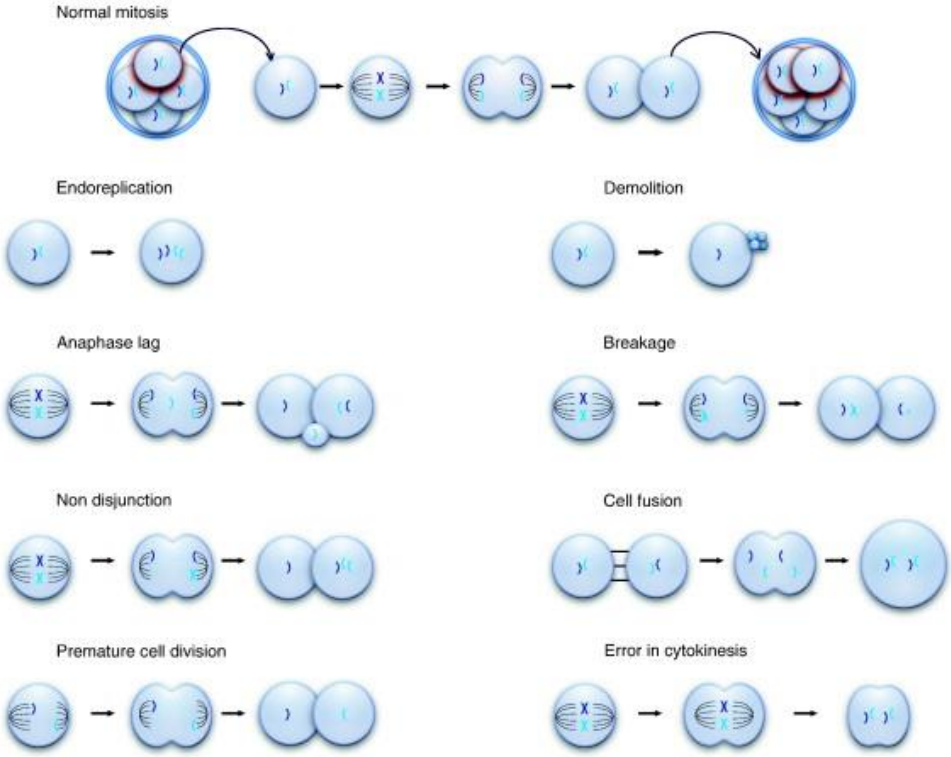
The failure of most IVF cycles is attributed to the transfer of aneuploid embryos (Tiegs et al. 2021). Concurrently, it is known that the proportion of aneuploid embryos significantly increases with the egg provider's age, such that by the age 42 approximately 75% of blastocysts are aneuploid (Esteves et al. 2019). The ploidy status of blastocysts is determined by preimplantation genetic testing for aneuploidy (PGT-A). In this procedure, several cells are taken from the trophectoderm of the developing embryo and sent for genetic analysis, more commonly using next generation sequencing (NGS). The blastocysts are cryopreserved awaiting the PGT-A results, with the hopeful transfer of a euploid embryo in a frozen embryo transfer cycle. Thus, many IVF professionals advocate the use of PGT-A in older age groups. It must also be considered that PGT-A avoids the transfer of an embryo which results in a child affected by a chromosomal condition and potentially reduces miscarriage risk (Neal et al. 2018).

Figure 2: Definitions for the dynamic monitoring of human embryo development. See pages 7-11 for definitions *(used with permission from Ciray et al., 2014).*



Aneuploidy is mostly derived from mitotic errors in human preimplantation embryos, particularly in the first cleavage after fertilisation (Mantikou et al. 2012). This is often owing to anaphase lagging or the failure of the chromatid to connect to the spindle apparatus and thus not being included in the nucleus of daughter cells. It may also be due to non-disjunction where the sister chromatids fail to separate properly during mitosis. These mechanisms and others are outlined in Figure 3.

Figure 3: Mechanisms of mitotic aneuploidy (used with permission from Mantikou et al. 2012)



Despite the potential benefits described, the use of PGT-A is a controversial topic within reproductive medicine. The regulatory body in the UK, the Human Fertilisation and Embryology Authority (HFEA) updated its classification of IVF 'add-ons' in October 2023 (HFEA 2023). Previously this stated that there is 'no evidence that PGT-A is effective', however this has now been updated; stating that PGT-A reduces the chance of miscarriage for most fertility patients, they state there is insufficient evidence to determine whether PGT-A improves the chances of having a baby for older patients, however, for most fertility patients it does not increase the chance of a live birth. This is due to no randomised controlled trial showing an improvement in the cumulative live birth rate (Cornelisse et al. 2020). Critics would argue that the outcome measured

should instead be live birth rate per embryo transfer and that existing study samples do not include participants who are most likely to benefit (Griffin 2022). Despite a lack of trial evidence, there are over 100 retrospective studies demonstrating a higher live birth or clinical pregnancy rate per embryo transfer and potentially a shorter time to pregnancy for those 37 years of age and older (Bhatt et al. 2021; Neal et al. 2018). Not least, the HFEA data itself, when reanalysed demonstrates improved clinical outcomes per embryo transfer, particularly in patients older than 35 (Sanders et al. 2021). Controversies aside, it is clear that the use of PGT-A is increasing, particularly in the United States, therefore finding less invasive, more cost effective options through technologies, such as morphokinetic algorithms, has been a focus of many research groups.

Traditionally, morphokinetic models used a basic hierarchical structure, however, over the last decade and with the increasing availability of data, the algorithms used have become more complex. In particular, artificial intelligence (AI) has gained traction in its ability to predict clinical outcomes using routinely obtained information, such as patient attributes and even blastocyst image analysis. AI is a term that describes the ability of machines to mimic human decision making. Machine learning is a subset of this technology that learns to process data without explicit programming. Deep learning is a further subset of machine learning that utilises artificial neural networks (ANN) which simulate the architecture of neurons in the human brain. It can send information forwards and backwards in order to compute a decision via multiple connections. Such methods can integrate a ‘black box’ approach whereby the system arrives at a decision but there is no explanation or understanding about how it arrived there. This results in

these systems being less interpretable than traditional statistics, such as logistic regression. In this thesis we will examine whether such approaches outperform more traditional methods when creating morphokinetic algorithms for ploidy prediction.

There are several mechanisms by which morphokinetics have biological plausibility to risk stratify embryos for ploidy status. The first consideration is the differential mitochondrial content that has been identified during embryo development (Ho et al. 2018). Aneuploid embryos have been found to have a lower mitochondrial content, thus possible accounting for abnormal or slower cell divisions in aneuploid embryos as detected by a TLS (Ho et al. 2018). Second, the delays we see in morphokinetics may be related to the activation of error detection and DNA repair mechanisms by the developing embryo (Coticchio et al. 2021). he immediate consequences of aneuploidy on cell physiology have largely been studied in cancer cells rather than human pre-implantation embryos. In particular, it has been identified that aneuploidy can lead to increased DNA damage, condensation defects, replication stress, and most importantly cell cycle delays (Andriani et al. 2016; Santaguida et al. 2017; Williams et al. 2008). Third, it has also been reported that embryos with a better prognosis as determined by a morphokinetic algorithm have improved processing of metabolites such as glucose and amino acids (Ferrick, Lee, and Gardner 2020). This may be because the changes to chromosome copy number lead to altered gene expression, proteome and metabolism, a phenomenon termed aneuploidy-stress (Zhu et al. 2018). These changes to the metabolomic profile of aneuploid blastocysts could therefore having implications for the speed of cytokinesis.

The detailed basic science leading to the downstream phenomena seen on a TLS are out of the realms of this PhD. Instead, the following chapters will report on associations between aneuploidy and morphokinetics. Furthermore, whether morphokinetic ploidy prediction models translate into improved clinical outcomes for the patient. This body of work represents a gap in the literature since no systematic review and meta-analysis has explored the question investigated herein. While there are several cohort studies describing the development of ploidy predictive models based on morphokinetics, this was the first one of this size.

OBJECTIVES

There are four objectives within this PhD, thus, these will be addressed over the next four chapters:

1. To determine whether morphokinetic variables and morphological features are associated with the ploidy status of pre-implantation human embryos.
2. To compare the discriminative ability of different models created using four different machine learning algorithms for predicting ploidy status using morphokinetics.
3. To examine whether the risk scores derived from morphokinetic models are associated with live birth and miscarriage. Further to this, this chapter aims to discover whether clinical factors should be incorporated into morphokinetic embryo selection models.
4. Finally, and arguably most important, this chapter aims to determine whether morphokinetic models are better at prioritising euploid embryos for transfer over selection by a senior embryologist.

CHAPTER 2: MORPHOLOGICAL AND MORPHOKINETIC ASSOCIATIONS WITH ANEUPLOID: A SYSTEMATIC REVIEW AND META-ANALYSIS

Thomas Bamford, Amy Barrie, Sue Montgomery, Rima Dhillon-Smith, Alison Campbell, Christina Easter, Arri Coomarasamy

Manuscript published in Human Reproduction Update, pp.1-31 on 25th March 2022

https://doi.org/10.1093/humupd/dmac022

My role in this publication is as follows: I devised the areas to be search and created the search terms, carried out the literature search, collected the data, undertook the meta-analysis and wrote the manuscript.

Human Reproduction Update, pp. 1–31, 2022
https://doi.org/10.1093/humupd/dmac022

human reproduction update

Morphological and morphokinetic associations with aneuploidy: a systematic review and meta-analysis

Thomas Bamford [1,*], Amy Barrie [2], Sue Montgomery [1], Rima Dhillon-Smith [3], Alison Campbell [4], Christina Easter [3], and Arri Coomarasamy [3]

[1]CARE Fertility Manchester, Manchester, UK [2]CARE Fertility Chester, Chester, UK [3]Tommy's National Centre for Miscarriage Research, Institute of Metabolism and Systems Research, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, UK [4]CARE Fertility, John Webster House, Nottingham, UK

*Correspondence address. CARE Fertility Manchester, 109-112 Daisy Bank Rd, Manchester M14 5QH. UK. E-mail: thomasbamford@nhs.net
https://orcid.org/0000-0001-9048-3912

Submitted on December 1, 2021; resubmitted on February 10, 2022; editorial decision on March 25, 2022

TABLE OF CONTENTS

BACKGROUND: A time lapse system (TLS) is utilized in some fertility clinics with the aim of predicting embryo viability and chance of live birth during IVF. It has been hypothesized that aneuploid embryos display altered morphokinetics as a consequence of their abnormal chromosome complement. Since aneuploidy is one of the fundamental reasons for IVF failure and miscarriage, attention has focused on utilizing morphokinetics to develop models to non-invasively risk stratify embryos for ploidy status. This could avoid or reduce the costs associated with pre-implantation genetic testing for aneuploidy (PGT-A). Furthermore, TLS have provided an understanding of the true prevalence of other dysmorphisms. Hypothetically, the incorporation of morphological features into a model could act synergistically, improving a model's discriminative ability to predict ploidy status.

OBJECTIVE AND RATIONALE: The aim of this systematic review and meta-analysis was to investigate associations between ploidy status and morphokinetic or morphological features commonly denoted on a TLS. This will determine the feasibility of a prediction model for euploidy and summarize the most useful prognostic markers to be included in model development.

SEARCH METHODS: Five separate searches were conducted in Medline, Embase, PubMed and Cinahl from inception to 1 July 2021. Search terms and word variants included, among others, PGT-A, ploidy, morphokinetics and time lapse, and the latter were successively substituted for the following morphological parameters: fragmentation, multinucleation, abnormal cleavage and contraction. Studies were limited to human studies.

OUTCOMES: Overall, 58 studies were included incorporating over 40 000 embryos. All except one study had a moderate risk of bias in at least one domain when assessed by the quality in prognostic studies tool. Ten morphokinetic variables were significantly delayed in aneuploid embryos. When excluding studies using less reliable genetic technologies, the most notable variables were: time to eight cells (t8, 1.13 h, 95% CI: 0.21–2.05; three studies; n = 742; I^2 = 0%), t9 (2.27 h, 95% CI: 0.5–4.03; two studies; n = 671; I^2 = 33%), time to formation of a full blastocyst (tB, 1.99 h, 95% CI 0.15-3.81; four studies; n = 1640; I^2 = 76%) and time to expanded blastocyst (tEB, 2.35 h, 95% CI: 0.06–4.63; four studies; n = 1640; I^2 = 83%). There is potentially some prognostic potential in the degree of fragmentation, multinucleation persisting to the four-cell stage and frequency of embryo contractions. Reverse cleavage was associated with euploidy in this meta-analysis; however, this article argues that these are likely spurious results requiring further investigation. There was no association with direct unequal cleavage in an embryo that progressed to a blastocyst, or with multinucleation assessed on Day 2 or at the two-cell stage. However, owing to heterogeneous results and poor-quality evidence, associations between these morphological components needs to be investigated further before conclusions can be reliably drawn.

WIDER IMPLICATIONS: This first systematic review and meta-analysis of morphological and morphokinetic associations with ploidy status demonstrates the most useful morphokinetic variables, namely t8, t9 and tEB to be included in future model development. There is considerable variability within aneuploid and euploid embryos making definitively classifying them impossible; however, it is feasible that embryos could be prioritized for biopsy. Furthermore, these results support the mechanism by which algorithms for live birth may have predictive ability, suggesting aneuploidy causes delayed cytokinesis. We highlight significant heterogeneity in our results secondary to local conditions and diverse patient populations, therefore calling for future models to be robustly developed and tested in-house. If successful, such a model would constitute a meaningful breakthrough when accessing PGT-A is unsuitable for couples.

Key words: time-lapse / morphokinetics / ploidy / model / fragmentation / multinucleation / abnormal cleavage / contraction

Introduction

Pre-implantation embryo selection has historically relied upon morphological assessment using increasingly contested consensus guidelines (Gardner, 1999; Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology, 2011; Gardner and Balaban, 2016; Kemper et al., 2021). Despite significant improvements since the inception of assisted reproduction, the average live birth rate in the UK remains low, at 32% per embryo transfer (for women <35 years) (HFEA, 2021a). When one also considers the drive for single embryo transfers, advancing maternal age and higher associated aneuploid rates, the need for more advanced methods for assessing embryo viability is paramount.

A time lapse system (TLS) offers several advantages over static, basic morphological observations. This enclosed incubation system reduces the need to remove embryos from optimum atmospheric culture conditions by taking microscopic, multiplanar images at regular intervals. The retrospective analysis of these images allows the annotations of an embryo's developmental milestones (i.e. morphokinetics) to be compared to outcome variables, such as live birth or ploidy status. This allows embryos to be selected that display specific development patterns achieved at fixed times of development, for example blastocyst formation at 116 h; usually recorded as hours post insemination (hpi). Unfortunately, due to poor-quality evidence, a Cochrane review was unable to conclude whether the use of a TLS increased live birth rates (Armstrong et al., 2019). In contrast, several large studies and randomized trials have reported improvements, therefore a TLS has become commonplace in many IVF laboratories worldwide (Pribenszky et al., 2017). A summary of definitions used for morphokinetic annotations and other morphological features denoted on a TLS can be found in Table I.

Aneuploidy is a major cause of implantation failure and miscarriage; however, there are barriers to accessing genetic testing. Aneuploidy arises from errors during mitosis or meiosis, such as non-disjunction. This increases with maternal age and therefore coincides with rapidly declining success rates of IVF treatment in older women. For instance, in women under 35 years, an average aneuploidy rate of 30–50% has been reported, increasing to 80% in women aged 42 years or older (Ata et al., 2012; Franasiak et al., 2014). Modern methods for pre-implantation genetic testing for aneuploidy (PGT-A) provide an

Table I Definitions of morphokinetic variables and morphological features in the human embryo.

Morphokinetics	The behaviour of an embryo as it develops, monitored using a time lapse system allowing embryologists to record 'developmental milestones' or morphokinetic variables retrospectively.
tPB2	Time to appearance of second polar body
tPNf	Time when both pronuclei have faded
tn	Time from insemination to completed division to n cells
tSC	Time from insemination to the start of compaction (when the first cells of the embryo join)
tM	Time from insemination to the formation of the morula, where all cells have undergone compaction and cell boundaries are unclear
tSB	Time from insemination to the start of blastulation when the first signs of cavitation are visible
tB	Time from insemination to the formation of a full blastocyst; when the blastocoele filled the embryo with <10% in diameter
tEB	Time from insemination to expanded blastocyst; when the blastocyst had increased in diameter by more than 30% and the zona pellucida starts to thin
tHB	Time from insemination to hatching blastocyst, trophectoderm herniation through the zona pellucida is observed
cc2	Time of the second cell cycle (t_3–t_4), from two to three cells
cc3	Time of the third cell cycle (t_5–t_3), from two to three cells
S2	Time of synchrony of second cell cycle (t_4–t_3) from two to four cells
S3	Time of synchrony of third cell cycle (t_8–t_5) from four to eight cells
Blastulation	Time of blastulation, from start of blastulation to formation of a full blastocyst (t_B–t_{SB})
Fragmentation	The presence of anucleated structures of blastomeric origin, degree of fragmentation is expressed as a percentage of total cytoplasmic volume (ESHRE 2000)
Abnormal cleavage (AC)	When the embryo does not display normal cytokinesis or division patterns one would normally expect. This can be broken down into different types (see below) (Barrie *et al.*, 2017)
Direct uneven cleavage (DUC)	Division of one blastomere into three instead of two daughter cells (Barrie *et al.*, 2017) DUC 1—abnormal division after syngamy resulting in 3–4 blastomeres DUC 2—abnormal cleavage at the two-cell stage resulting in five or six blastomeres DUC 3—abnormal cleavage at the four-cell stage resulting in nine blastomeres or more (Zhan *et al.*, 2016)
Reverse cleavage (RC)	Results from blastomere fusion (Barrie *et al.*, 2017)
Embryo contraction	Defined as a spontaneous separation of the pellucid zone and trophectoderm that occurs from the blastocyst stage (Marcos *et al.*, 2015)
Multinucleation	Defined as the presence of more than one nucleus in at least one blastomere of the embryo, this can be assessed at the two or four cell stage or alternatively on Day 2 or Day 4.

Table reproduced and adapted with permission from author AC's article (Campbell *et al.*, 2013a).

accurate assessment of embryo chromosome complement using biopsy techniques in the majority of cases (Munné *et al.*, 2017, 2019). For some patients, however, this technology may be inaccessible because it is prohibited by legislation, or they may deem it ethically inappropriate. They may also not have embryos suitable for biopsy. Moreover, PGT-A can cost over £3000 in the UK and in the USA it can be as high as $12 000, further limiting accessibility (Theobald *et al.*, 2020). It is therefore not surprising that researchers have begun investigating methods to non-invasively detect aneuploidy.

It has been hypothesized that the morphokinetics of aneuploid embryos are delayed in comparison to euploid counterparts (Davies *et al.*, 2012; Campbell *et al.*, 2013a). Physiologically, this may be due to complex biochemical processes that occur when errors have been detected by the developing embryo (Coticchio *et al.*, 2021a). This results in slower cell division and is possibly a reason for the higher mitochondrial content seen in aneuploid embryos (Campbell *et al.*, 2013a; Ho *et al.*, 2018). This has led to the development of several models using PGT-A and morphokinetic data aiming to risk-stratify embryos for euploidy (Campbell *et al.*, 2013a; Basile *et al.*, 2014; Chawla *et al.*, 2015; Del Carmen Nogales *et al.*, 2017; Desai *et al.*, 2018; Mumusoglu *et al.*, 2017). At CARE Fertility, a sophisticated time-lapse embryo selection model, 'CAREmaps®', has been successfully developed that can predict an individual embryo's chance of resulting in a live birth. This was developed using a database of over 6000 transferred blastocysts with known live birth outcome data and has been shown to improve embryo selection (Fishel *et al.*, 2018). Similar embryo selection algorithms have been developed by a variety of clinics internationally; some are commercially available (Petersen *et al.*, 2016). It remains unknown why embryos with higher scores should have better predicted outcomes; it would be sensible to hypothesize that the aetiology lies within delayed development as a sequela of chromosomal abnormalities. It would therefore prompt the assumption that if a TLS can identify embryos with the highest chance of live birth, it could be instrumented to enhance euploid embryo selection. This hypothesis is also supported by a recent meta-analysis that showed that the use of a TLS was associated with lower early miscarriage rates compared to traditional morphological assessment (Pribenszky *et al.*, 2017). Other theories have also been suggested including partial compaction with or without cell extrusion or exclusion causing delayed cyto or karyokinesis, abnormal fertilization, BMI,

embryo sex, a failure of the embryo to undergo check points and DNA repair mechanisms (Bronet et al., 2015; Leary et al., 2015; Coticchio et al., 2021a,b). It may therefore be feasible to utilize morphokinetics as a screening tool for ploidy status if this hypothesis becomes established by evidence.

The ability of morphokinetic models to predict ploidy status remains controversial and wide disparities exist in the morphokinetic events included in such models (Campbell et al., 2013b; Basile et al., 2014; Kramer et al., 2014). This may be due to significant heterogeneity in study design and sample populations. For instance, the following have all been associated with altered morphokinetics: age, smoking status, biopsy techniques, stimulation protocols, insemination methods and culture conditions (Lemmen et al., 2008; Ciray et al., 2012; Muñoz et al., 2013; Bellver et al., 2013; Fréour et al., 2013; Kirkegaard et al., 2013b).

Several morphological observations can be observed in greater detail when using a TLS, although historically there is limited correlation reported between ploidy status and these qualitative aspects (Magli et al., 2007; Capalbo et al., 2014; Minasi et al., 2016; Munné et al., 2017). In fact, several authors have identified that it is possible for aneuploid embryos to achieve good morphology scores (Munné, 2006; Alfarawati et al., 2011; Fragouli et al., 2014). Nonetheless, it must be taken into consideration that most studies investigating associations between morphology and ploidy status were undertaken using standard morphology assessments and not using a TLS. This results in an inability to identify dynamic changes occurring between check points. Furthermore, many of these studies utilized older, less reliable techniques such as fluorescence *in situ* hybridization (FISH) and blastomere biopsy. This results in a higher chance of misclassifying mosaics or failing to detect aneuploidy due to the limited number of probes used (Fragouli and Wells, 2011). We aim to investigate the association of various morphological components commonly observed on a TLS with ploidy status.

The first variable to be explored is fragmentation. Fragmentation is often considered during embryo selection owing to associations with embryo viability, but it remains one of the most enigmatic features identified in early development (Edwards et al., 1984; Puissant et al., 1987). Origins of these anucleated structures have been correlated with many factors including culture conditions, poor-quality oocytes or spermatozoon, increased maternal age, oxidative stress and aneuploidy (Munne and Cohen, 1998; Delimitreva et al., 2005; Magli et al., 2007; Fujimoto et al., 2011; Kim et al., 2018). It has even been associated with so called 'self-correction' mechanisms whereby an embryo extrudes sequestered chromosomes in order to become more genetically normal (Coticchio et al., 2021a). Considering that the causation is poorly understood, association with ploidy status will be explored further in this review.

The second factor to be investigated is abnormal cleavage, the occurrence of which has become more apparent through a TLS yet causality remains unproven (Athayde Wirka et al., 2014; Zhan et al., 2016). The prevalence of these atypical cell divisions ranges from 4.4 to 26.1% and the implantation rates of these untested embryos has been found to be as low as 1.2–17% (Rubio et al., 2012; Barrie et al., 2017; Ozbek et al., 2021). There is, therefore, a tendency to deselect these embryos (Balakier et al., 2016; Hashimoto et al., 2016; Zhan et al., 2016; Desai et al., 2018). Previous theories for aetiology include multipolar spindles, surplus centrosomes, quality of spermatozoa and chromosome aberrations (Kalatova et al., 2015; Ozbek et al., 2021). Similarly, it has been speculated that abnormal cleavage may also be involved in the process of 'self-correction'. This is supported by the recent findings that abnormal cleavages are associated with partial compaction and the 'excluded phenotype' (Coticchio et al., 2021b). These excluded cells have also been shown to have a significantly higher abnormal chromosome content (Lagalla et al., 2017). We will assess the association between the most common types of abnormal cleavage and ploidy status: direct and reverse cleavage (Rubio et al., 2012; Liu et al., 2014).

Blastocyst contraction is the third feature to be examined that has been the focus of only a handful of studies. Physiologically this occurs through the inflow of liquid through aquaporin water channels and outflow through weak tight junctions (Watson et al., 2004; Marcos et al., 2015). The reason for it remains largely unknown, and it has been suggested that this process may assist in embryo hatching and has been associated with lower implantation rates (Niimura, 2003; Marcos et al., 2015; Bodri et al., 2016). Hypothetically, this may be secondary to aneuploidy, therefore this will be investigated in this review.

Finally, multinucleation has been associated with poorer implantation outcomes and possibly aneuploidy (Kligman et al., 1996; Royen et al., 2003). This dysmorphism has been hypothesized to be the result of errors in nuclear replication without cytokinesis, nuclear fragmentation or defective DNA packaging and migration during anaphase (Pickering et al., 1995). It is therefore possible that this could be linked to aneuploidy as a consequence of errors occurring in chromosome segregation.

The aim of this systematic review and meta-analysis is to determine the most reliable morphokinetic prognostic factors for future model development and investigate associations between morphology and ploidy status. Specifically, the degree of fragmentation, presence of direct and reverse cleavage, blastocyst contractions and multinucleation will be investigated in association with chromosomal status. Incorporating these morphological parameters may improve the discrimination of a morphokinetic model with regards to ploidy.

Methods

Registration

This review was prospectively registered with PROSPERO (ID number: CRD42021260795).

Data sources and search strategy

Five separate literature searches were conducted for potential prognostic factors and their associations with aneuploidy in concordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Moher et al., 2009). Electronic searches were conducted in MEDLINE, PubMed, EMBASE and CINAHL (from inception to 1 July 2021). Searches were conducted using the following MeSH key terms and word variants: 'pre-implantation genetic testing for aneuploidy (PGT-A)', OR 'pre-implantation genetic screening (PGS)', OR 'ploidy', OR 'aneuploid' AND 'morphokinetics', OR 'time-lapse'. For the four subsequent searches 'morphokinetics' and 'time-lapse' were successively substituted for: 'fragmentation',

'multinucleation', 'abnormal cleavage' and 'blastocyst contraction'. Similarly, word variants for each were included, such as 'trichotomous mitosis' for 'direct cleavage'.

Eligibility criteria

Studies were limited to human studies and included if the primary or secondary outcome was the ploidy status of biopsied embryos in relation to the presence of any of the prognostic factors under investigation. No language restrictions were applied. Manuscripts on mosaicism were included if they also provided data on aneuploid and euploid embryos. Exclusions include: polar body biopsy, those reporting clinical outcomes only, where the outcome was translocations not aneuploidy, those that focussed on a subset of embryos with a particular morphological feature (such as abnormal cleavage in multinucleated embryos) or from a subgroup of patients (for example, endometriosis). The blastocyst contraction literature search aimed to determine association of embryo contraction kinetics (number or frequency), therefore studies investigating the rate or volume of expansion in relation to ploidy were excluded. Similarly, authors that correlated blastocyst expansion grading or morphology scores with ploidy status but not the individual prognostic factors being tested were excluded. Validation studies for a prognostic model already developed were excluded from the meta-analysis.

Study selection

Two reviewers initially screened all titles and abstracts independently for eligibility (T.B. and A.B.), and full length articles were then obtained and scrutinized. Any disputes were resolved by discussion with a third reviewer (S.M.). Bibliographies of all relevant articles and review articles excluded were manually searched. Where more than 10 original articles met eligibility criteria, conference abstracts were subsequently excluded from the search. Otherwise, they were included due to a scarcity of published peer reviewed reports. Authors of all conference abstracts were contacted for additional information to assist with study selection, data extraction and quality assessments. Authors of original articles were contacted for further information where data presented were suboptimal. If data were not obtained or in a usable format, it was excluded from the meta-analysis but included in the systematic review.

Data extraction and study outcomes

Outcome and prognostic factor data were extracted independently by two reviewers into tables (T.B. and A.B.). The primary outcome extracted was the prevalence of aneuploid and euploid embryos for each potential prognostic factor assessed. This included the mean or median time taken for both aneuploid and euploid embryos to achieve each morphokinetic variable. Data were also collected for: overall aneuploid rate, study design, primary outcome measured, number of patients and embryos included, TLS assessment period, PGT-A technique (including stage and type of biopsy), atmospheric culture conditions, infertility diagnosis and indication for PGT-A. Additionally, details of any model development, including attempts at model discrimination, calibration and validation, were recorded. Importantly, we collected data on potential study participant factors that could act as confounders including age, BMI and stimulation drugs used. A recently published article by Barrie *et al.* (2021) described how age and BMI are the most important factors to control for in morphokinetic studies. Data were extracted only on those embryos with PGT-A results available.

Risk of bias and quality assessment

All articles meeting the selection criteria were quality assessed using the Quality in Prognosis Studies tool (QUIPS) (Grooten *et al.*, 2019). It moves away from quantitative analysis of quality but rates the risk of bias in six domains (participation, attrition, prognostic factor measurement, outcome measurement, study confounding and statistical analysis) as low, medium or high risk of bias (Higgins *et al.*, 2021). The tool has been modified for use in this systematic review: an example can be found in Supplementary Table SI, including a summary of the bias domains and the criteria used to grade each category. Several items were removed from our adapted version of the tool. First, the 'adequate participation', 'drop out' and 'attempt to collect information on participants who dropped out' prompting items were removed because they were less relevant to the study of embryos as research focuses on the retrospective analysis of existing PGT-A data sets. The original tool included a prompter within the confounding domain asking the reviewer to determine if the method used to measure confounding was reliable. This was removed because confounders for morphokinetics include readily available demographic data and standardized dosages.

There were a number of important factors to consider when undertaking the quality assessment. First, if a particular study did not report on the proportion of embryos without PGT-A results, they were categorized as having a moderate risk of bias and if it was >5%, they were deemed to have a high risk of bias. Using modern methods, most genetic companies would now estimate that this occurs in up to ~2% of samples and this has been replicated in recent studies (Fiorentino *et al.*, 2014; Neal *et al.*, 2019; Tiegs *et al.*, 2021). A significantly high proportion may lead to uncertainty regarding the validity of a study's conclusions and biopsy techniques. Second, if FISH was one of the genetic platforms used by a study, it was considered a high risk of bias owing to the inaccuracies of this technique. Finally, due to the risk of inter-observer variability in morphological assessments of embryos, if there were no methods to account for internal validity, then a publication was assessed as a moderate risk of bias. Similarly, if multinucleation was assessed as part of standard morphology assessment rather than the continuous observations enabled by a TLS, it was graded as a moderate risk of bias.

As per the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines, publication bias was not assessed as <10 studies were included for each prognostic factor analysed, rendering the interpretation of funnel plots unreliable (Schünemann *et al.*, 2013). The quality of reporting was assessed using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist according to published criteria (von Elm *et al.*, 2007).

Data synthesis and analysis

Morphokinetics of aneuploid and euploid embryos were compared using a weighted mean difference analysis in concordance with the Meta-Analysis of Observational Studies in Epidemiology recommendations (Stroup, 2000). Where studies did not provide a SD value, it was

calculated from the 95% CI if the sample size was >100 (using methods recommended by the Cochrane Handbook) (Higgins et al., 2021). Alternatively, data from studies describing only medians were summarized graphically by prognostic factor analysis using the interquartile range as a measure of dispersion.

The relationship between the prevalence of aneuploid embryos and percentage fragmentation has been presented on a line graph. Abnormal cleavage, embryo contraction and multinucleation data were dichotomized and meta-analysed with forest plots and corresponding calculated relative risks (RR). The results for abnormal cleavage were pooled and analyses were conducted for each type of abnormal cleavage to determine their relative contribution.

For all meta-analysed variables, heterogeneity was assessed using the I^2 statistic, whereby a result >50% was indicative of considerable heterogeneity. All analyses were undertaken using a random effects model by the Mantel–Haenszel method (Mantel and Haenszel, 1959; DerSimonian and Laird, 1986) using Review Manager (RevMan), version 5.4.

Sensitivity analysis

Sensitivity analyses were conducted based on the quality assessment derived from QUIPS. The use of random versus fixed effect models were also compared. Analyses will be restricted by excluding studies with a high risk of bias in any domain, an approach also taken by other authors using QUIPS (Taylor-Rowan et al., 2021).

Subgroup analyses

A subgroup analysis was performed by excluding studies using FISH, blastomere or Day 3 biopsy to produce a more reliable effect estimate. During the last decade, trophectoderm biopsy at the blastocyst stage has become the preferred method of testing; fewer embryos have mosaic results and there is less risk of damage and diminishing the live birth rates (Tarín et al., 1992; Los et al., 2004; Staessen et al., 2004; Cohen et al., 2007; Goossens et al., 2008). FISH also has several limitations, primarily, the impossibility to screen all chromosomes and the risk of misdiagnosis is significant when multiple probes are used (Ruangvutilert et al., 2000; DeUgarte et al., 2008; Scriven and Bossuyt, 2010). These studies were not excluded as part of the screening process as they may still provide valuable prognostic information. None of the studies included in the prognostic factor graphs used FISH or blastomere biopsy.

Results

Search results

A total of 1557 studies were identified from the initial searches; 137 duplicates were removed, 1267 abstracts were screened, of which 123 were selected as being potentially relevant and 58 met selection criteria after screening of the full texts. Overall, 65 studies were excluded for the following main reasons: a different prognostic factor being investigated (n = 15), describing an alternative outcome to ploidy status (n = 11), measuring clinical outcomes only (n = 10), or the study population was a subset of embryos with a particular characteristic (n = 7). Figures 1–5 display the study screening process for each search and all exclusions are summarized in Supplementary Table SII (Moher et al., 2009). In total, 26 authors were contacted to attempt to identify missing information from their publication (Supplementary Table SIII). Two studies are 'awaiting classification' due to unanswered correspondence with the authors: this was intended to confirm whether their abstracts included different embryos than the later published articles (Lagalla, 2015; Desai and Rambhia, 2016). Responses were not received, therefore the publications with the most data were included in this systematic review (Lagalla et al., 2017; Desai et al., 2018). Responses were not received from 19 authors in total; this did not result in exclusion from the systematic review for any of these studies but exclusion from the quantitative analysis in 10. The remaining were contacted for supporting information only. In total, 58 studies were included in the narrative synthesis, 43 of which had results suitable for meta-analysis.

Characteristics of the included studies

Overall, 7004 embryos that underwent PGT-A from at least 1058 patients were included from 18 studies examining morphokinetics. Thirteen retrospective studies and one prospective cohort study provided morphokinetic data comparing euploid and aneuploid embryos, enabling their inclusion in the quantitative analysis (Chavez et al., 2012; Campbell et al., 2013a; Basile et al., 2014; Yang et al., 2014; Chawla et al., 2015; Rienzi et al., 2015; Minasi et al., 2016; Patel et al., 2016; Zhang et al., 2017; Desai et al., 2018; Kimelman et al., 2019; Lee et al., 2019; Martin et al., 2021; Mumusoglu et al., 2017). Three studies were excluded from meta-analysis because two were validation studies (Campbell et al., 2013a; Kramer et al., 2014) and one study presented data in an unusable format (Del Carmen Nogales et al., 2017). The included studies for morphokinetics were from eight different countries in total (USA, UK, Spain, Italy, India, Turkey, Taiwan and China). A summary of the characteristics of the included studies are reported in Tables II–VI.

Subsequently in this review, we have considered morphological associations of aneuploidy. First, fragmentation was assessed using data from 10 008 embryos from 1842 patients, extracted from five studies (Magli et al., 2001; Ziebe et al., 2003; Delimitreva et al., 2005; Magli et al., 2007; Minasi et al., 2016). The remaining studies provided no raw data for interpretation or when provided it was in an unusable format (Moayeri et al., 2008; Chavez et al., 2012; Vera-Rodriguez et al., 2015). All were retrospective cohort studies apart from one publication, which was a consecutive case series (Minasi et al., 2016).

Second, abnormal cleavage was assessed in relation to ploidy status, and this included 4788 embryos from 1100 patients from 10 retrospective cohort studies (Campbell et al., 2013a; Rienzi et al., 2013; Vera-Rodriguez et al., 2015; Zhan et al., 2016; Lagalla et al., 2017; Zhang et al., 2017; Desai et al., 2018; Ho et al., 2018; McCoy et al., 2018; Ozbek et al., 2021). One study could not be included in the meta-analysis due to the limited provision of data (Davies et al., 2012).

Third, two cohort studies were meta-analysed to describe the relationship between embryo contraction and chromosome aberrations using data from 1647 embryos from 460 patients (Vinals Gonzalez et al., 2018; Gazzo et al., 2020).

Finally, the presence of multinucleation was assessed in 18 676 embryos from 1227 patients. Thirteen studies were included in the meta-analysis (Kligman et al., 1996; Magli et al., 2001; Agerholm et al., 2008;

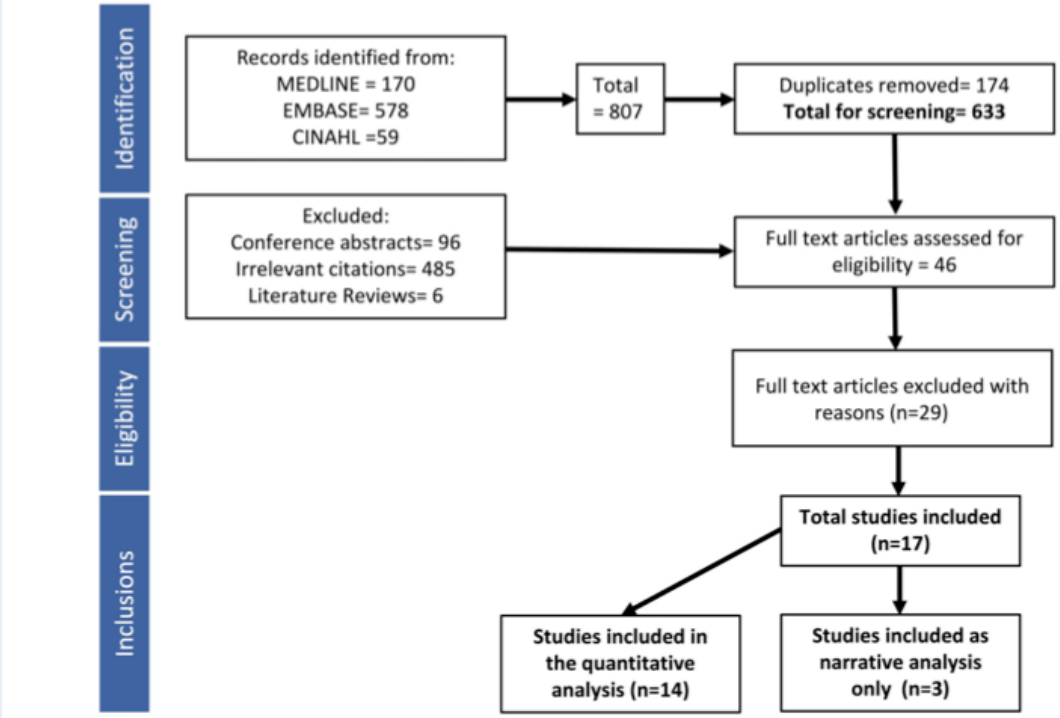


Figure 1. Flow diagram for study selection process for human embryo morphokinetics search.

Ambroggio *et al.*, 2011; Campbell *et al.*, 2013a; Mazur, 2013; Munoz *et al.*, 2014; Bayram, 2015; Balakier *et al.*, 2016; Hashimoto *et al.*, 2016; Zhang *et al.*, 2017; Desai *et al.*, 2018; Lee *et al.*, 2019) and seven studies were included only as part of the narrative review (Scott *et al.*, 2010; Davies *et al.*, 2012; Melzer *et al.*, 2013; Yilmaz *et al.*, 2014; Li *et al.*, 2015; Goodman *et al.*, 2016; Del Carmen Nogales *et al.*, 2017). Eleven are cohort studies and the remaining nine are conference abstracts (Davies *et al.*, 2012; Mazur, 2013; Melzer *et al.*, 2013; Munoz *et al.*, 2014; Bayram, 2015; Li *et al.*, 2015; Goodman *et al.*, 2016; Del Carmen Nogales *et al.*, 2017). The included manuscripts considering morphological prognostic factors were published from a broad range of countries (Tables II–VI).

Risk of bias and quality assessment results

Overall, the quality assessment of the eligible studies demonstrated a moderate risk of bias, whereby all but one study was scored with a moderate risk of bias in at least one domain. In total, only 17 out of 58 studies (29%) appropriately addressed confounding. Similarly, few authors adequately described participant characteristics or the selection criteria used (n = 18/58, 31%). However, there was a low risk of bias for 'prognostic factor measurement' in most studies (n = 45/58, 78%). The remaining studies had a moderate risk of bias within this category due to: unclear definitions of the prognostic factors (n = 3), a lack of internal validation for the assessment of the morphological components (n = 5), the use of standard morphology assessment at specific time points rather than the use of a TLS (n = 4) or multiple methods used for prognostic factor measurement on the same cohort (n = 1). Twenty-three studies were considered a moderate risk of bias because they did not disclose the proportion of embryos with PGT-A results unavailable, and five studies had a high risk of bias since this was >5%. Furthermore, 11 studies were categorized as high risk of bias for the use of FISH and one study for 'statistics and reporting' as we consider their conclusions and results to be erroneous given the data presented (Davies *et al.*, 2012). Finally, 17 studies within the 'statistics and reporting' domain were graded as a moderate risk of bias for: limited presentation of analytical strategy or data (n = 10), poor modelling techniques and validation methods (n = 5) and inappropriate statistical techniques (n = 2). The results obtained using the QUIPS tool to determine the risk of bias are summarized in Supplementary Table SIV. The quality of reporting was assessed using STROBE (Supplementary Fig. S1).

Morphokinetics and ploidy

The following morphokinetic variables (Table I) were significantly delayed in aneuploid embryos: tPB2, t2, t4, t6, t7, t8, t9, tB, time to expanded blastocyst (tEB) and tHB. In contrast, tPNf, tM, tSB,

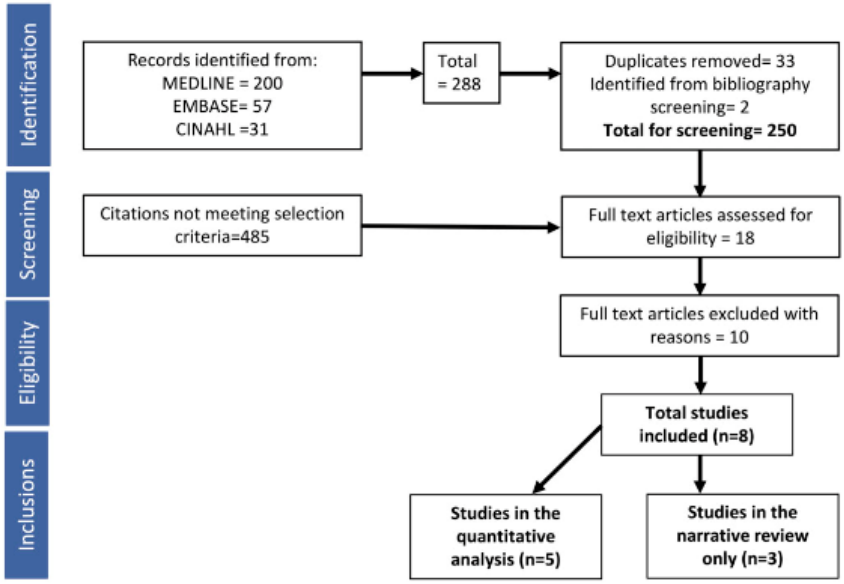


Figure 2. **Flow diagram for study selection process for human embryo fragmentation search.**

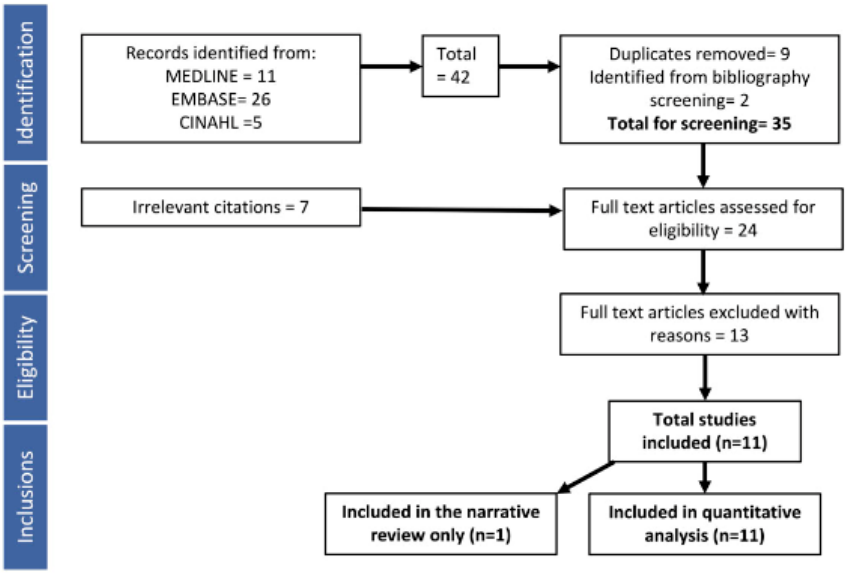


Figure 3. **Flow diagram for study selection process for human embryo abnormal cleavage search.**

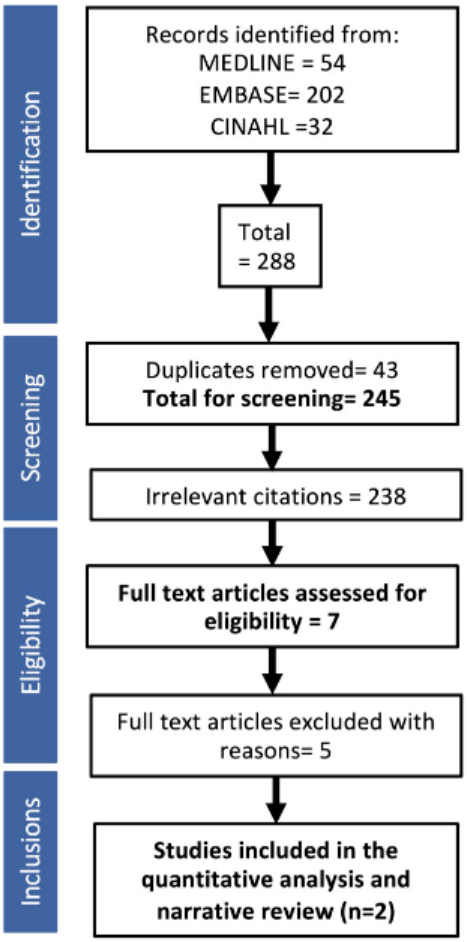


Figure 4. Flow diagram for study selection process for human embryo contraction search.

cc3, S2, S3 and t5-t2 had no prognostic ability (Supplementary Fig. S2). Interestingly, euploid embryos were significantly delayed for cc2; however, this finding, t6 and tHB were no longer statistically significant when studies using FISH and/or blastomere biopsy were excluded. Additionally, t3, t5 and tPNf demonstrated significant differences exclusively in the subgroup analysis. The variables tPNf, t2, t3, t4 and t5 were all delayed by up to 1 h in aneuploid embryos. The variables tPB2 and t7 were delayed by >1 h in aneuploid embryos; however, these results come from a subgroup analysis including only one study (1.3 h, 95% CI: 0.88–1.72 and 1.8 h, 95% CI: 0.34–3.26, respectively). The following variables were the most delayed in aneuploid embryos: t8 (1.13 h, 95% CI: 0.21–2.05; three studies; n = 742; $I^2 = 0\%$), t9 (2.27 h, 95% CI: 0.5–4.03; two studies; n = 671; $I^2 = 33\%$), tB (1.99 h, 95% CI: 0.15–3.81; four studies; n = 1640; $I^2 = 76\%$) and tEB (2.35 h, 95% CI: 0.06–4.63; four studies; n = 1640; $I^2 = 83\%$) (Fig. 6).

On visual inspection, these results were concordant with the prognostic factor graphs, apart from t8. Similarly, tSC, which was analysed solely by a prognostic factor graph, resulted in inconsistent differences. The only study excluded from the meta-analysis that was not a validation study analysed morphokinetics per chromosomal abnormality and found that complex embryos had shorter cleavage times (Del Carmen Nogales *et al.*, 2017). Finally, the sensitivity analyses did not change our conclusions with the exception of tSB that became significant using a fixed effects model (Supplementary Figs S3 and S4).

Fragmentation and ploidy

Fragmentation was associated with aneuploidy in six out of the eight included studies. The three most recent studies had the lowest risk of bias; two found no association and one found that a higher degree fragmentation was associated with aneuploidy

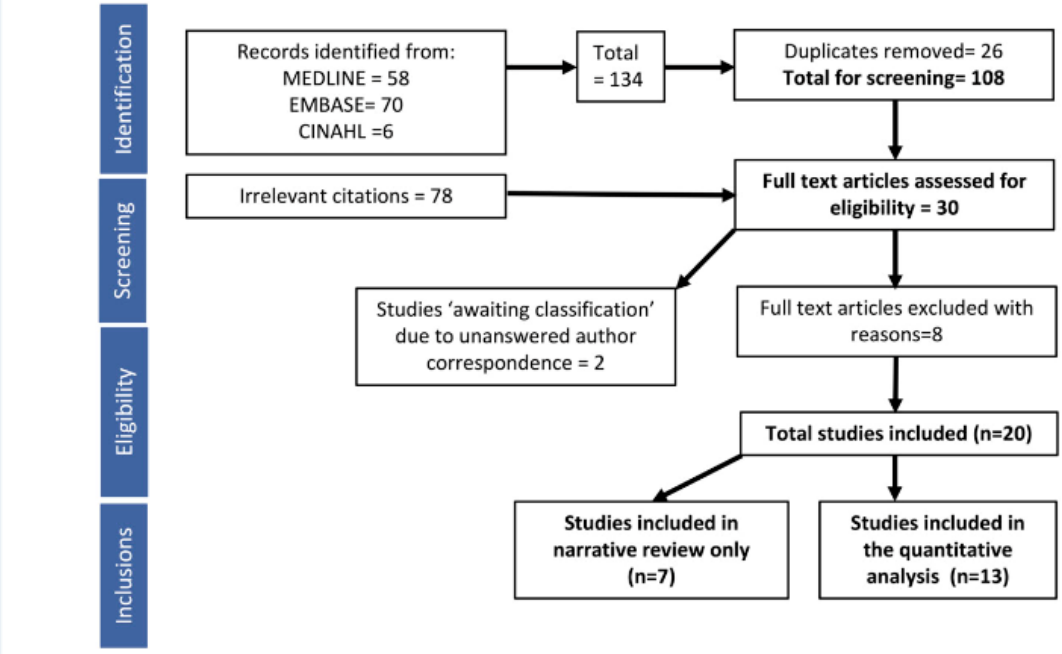


Figure 5. Flow diagram for study selection process for human embryo multinucleation search.

(Chavez *et al.*, 2012; Vera-Rodriguez *et al.*, 2015; Minasi *et al.*, 2016). Only four authors in total provided raw data that could be extracted into a line graph displaying a general trend of increasing prevalence of aneuploid embryos for increasing degrees of fragmentation (Supplementary Fig. S5).

Abnormal cleavage and ploidy

Pooled direct uneven cleavage, DUC1 and DUC2 (Table I), had no association with chromosomal normality (RR: 1.09, 95% CI: 0.83–1.44; RR: 1.26, 95% CI: 0.98–1.61; RR: 0.74, 95% CI: 0.26–2.1, respectively) (Supplementary Fig. S6). In contrast, reverse cleavage appears to provide some prognostic information specifically for euploidy (RR: 1.36, 95% CI: 1.14–1.63; five studies; n = 3053; $I^2 = 22\%$) (Fig. 7). There was a trend for more aneuploid embryos displaying DUC1 when studies using FISH and blastomere biopsy were excluded; however, this was not statistically significant (RR: 1.26, 95% CI: 0.98–1.61; five studies, n = 1917; $I^2 = 27\%$) (Supplementary Fig. S6). Only one study was not included in the meta-analysis owing to limited provision of data, and this concluded that embryos exhibiting DUC1 were more likely be aneuploid (57%, n = 21) versus euploid (30%, n = 44), $P = 0.01$) (Davies *et al.*, 2012). Our findings were unchanged in a sensitivity analysis excluding studies with the highest risk of bias (Supplementary Fig. S7). A sensitivity analysis using a fixed effect model resulted in DUC2 being significantly more prevalent in aneuploid embryos relative to euploid (Supplementary Fig. S8).

Contractions and ploidy

Two studies examined the association between the presence of contractions and ploidy status and found that this observation was significantly more likely to occur in aneuploid embryos (RR: 0.67, 95% CI: 0.48–0.96; two studies, n = 1626, $I^2 = 84\%$) (Fig. 8). These findings remained consistent in the sensitivity analysis (Supplementary Fig. S7).

Multinucleation and ploidy

No association with ploidy was found for embryos assessed on Day 2 or at the two-cell stage for multinucleation (RR; 0.69, 95% CI: 0.29–1.63, four studies, n = 3650, $I^2 = 0\%$; RR: 0.82, 95% CI: 0.64–1.04, seven studies, n = 2418, $I^2 = 47\%$, respectively); however, there may be prognostic potential in multinucleation persisting to the four-cell stage (RR: 0.52, 95% CI: 0.29–0.91; six studies, n = 1703, $I^2 = 82\%$) (Supplementary Fig. S9 and Fig. 9, respectively). This remains uncertain since the subgroup analysis was insignificant, albeit trending towards an increased prevalence in aneuploid embryos (RR: 0.56, 95% CI: 0.28–1.14; four studies, n = 1106, $I^2 = 88\%$). Furthermore, four-cell multinucleation was significantly associated with ploidy using a fixed rather than random effects model (Supplementary Fig. S8). Multinucleation on Day 2 and at the two-cell stage also had conflicting results in this sensitivity analysis; both were associated with aneuploidy but they remained insignificant in the subgroup analysis. Of the seven studies not included in the meta-analysis, two demonstrated association with ploidy when multinucleation was assessed during standard morphology assessments and one at the four-cell stage (Scott et al., 2010; Melzer *et al.*, 2013;

Table II Characteristics of included studies for morphokinetics and ploidy.

Morphokinetics and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factors-Method, assessment period, morphokinetic variables**	**Outcome measurement-PGT-A platform, stage and type of biopsy (aneuploidy rate)**	**Confounding factors and attempted adjustment yes (✓) or no (✗)**			**Main findings**	**Included in quantitative analysis (yes or no)**
						Mean age (SD)	**BMI (SD)**	**Stimulation**		
Chavez *et al.*, 2012, USA Retrospective Cohort, *Unknown*	Embryos donated for research, prognosis unknown	Unknown	53	Custom built TLS, 5 min, *Cc2, S2*	aCGH, Day 2 blastomere (75%)	33.5 ✗	Unknown ✗	✗	70% of aneuploid embryos exhibited parameters outside 'normal' values. Aneuploid embryos exhibit more diverse morphokinetic variables than euploid.	Yes
Campbell *et al.*, 2013a, UK, Retrospective Cohort, *May 2011 to July 2012*	IVF PGT-A cycles for: RIF, RM, SMF, PA, AMA	25	98	Embryoscope, 10 min, *tPNf, t2, t3, t5, t8, tSC, tM, tSB, t, tEB, tHB, cc2, cc3, s2, s3, Blastulation*	aCGH or SNP array, Day 5/6 Trophectoderm, (61%)	38.6 (3.6) ✗	Unknown ✗	✓ *No dose dependent differences in ploidy*	Aneuploid embryos were delayed at initiation of compaction and time to initiate and reach full blastulation.	Yes
Campbell *et al.*, 2013b, UK, Retrospective Cohort and Validation study, *April 2011 to December 2012*	IVF PGT-A cycles, indications unknown	69	88	Embryoscope, 20 min *Model validation study*	aCGH or SNP array, Day 5/6 Trophectoderm, (Unknown)	36.6 (5.1) ✗ *(Subsequently found Morphokinetics not affected by age)*	Unknown ✗	✗	Uses a model to demonstrate improvements in clinical outcomes but doesn't attempt to confirm the ability to improve the euploidy rate.	No Validation study
Yang *et al.*, 2014, USA, Prospective Cohort, *February to December 2012*	IVF PGT-A cycles, indications unknown	138	285	Embryoscope, 20 min, *t5, tSB, tB, tEB, cc2, S2*	aCGH or Whole Genome Amplification, Day 5 trophectoderm (51%)	36.6 (2.4) ✗ *(Inclusion criteria: all patients < 39 years of age)*	Unknown ✗	✗	No significant difference (although a slight delay) in morphokinetic variables in aneuploid vs. euploid embryos.	Yes
Basile *et al.*, 2014, Spain, Retrospective Cohort, *March 2012 to August 2012*	IVF PGT-A cycles for: RM (n = 40) and RIF (n = 37)	125	504	Embryoscope, 15 min, *tPN, tPNf, t2, t3, t4, t5, cc2, cc3, t5-t2*	aCGH, Day 3 blastomere (72%)	36.1 ✗	Unknown ✗	✗	t5–t2 was significantly different in aneuploid embryos. Suggests optimal ranges for t5 (47.2–58.2 h), cc3 (11/7–18.2 h) and t5-t2 (>20.5 h).	Yes
Kramer *et al.*, 2014, USA, Retrospective Cohort and Validation study, *Unknown*	IVF PGT-A cycles for: RM (n = 8), translocation (n = 1), family balancing (n = 2), AMA (n = 11), unexplained infertility (n = 5)	25	149	Embryoscope, 20 min, *Model validation study*	aCGH, Day 5 Trophectoderm (57%)	37.3 (3.9) ✗	Unknown ✗	✗	Tests the Campbell model. Failed to segregate euploid and aneuploid embryos. Time to compaction was the only variable to yield an AUC indicative of having predictive value (0.674).	No- Validation study and timed from syngamy not insemination

Continued

Table II Continued

Morphokinetics and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factors-Method, assessment period, morphokinetic variables**	**Outcome measurement-PGT-A platform, stage and type of biopsy (aneuploidy rate)**	**Confounding factors and attempted adjustment yes (✓) or no (✗)**			**Main findings**	**Included in quantitative analysis (yes or no)**
						Mean age (SD)	**BMI (SD)**	**Stimulation**		
Chawla *et al.*, 2015, USA, Retrospective Cohort, *May 2013 to May 2014*	IVF PGT-A cycles for: Sex selection	132	460	Embryoscope, 15min, *tPB2, tPN, tPNf, t2, t3, t4, t5, cc2, cc3, S2, t5-t2*	aCGH, Day 3 Blastomere (57%)	32.9 ✗	Unknown ✗	✗	tPNf, t2, t5, cc2, cc3, t5-t2 differed significantly in aneuploid embryos. Cc3 > 10 h and t5-t2 > 20 hours was optimal time for predicting normal embryos.	Yes
Rienzi *et al.*, 2015, Italy, Retrospective Cohort *December 2012 to December 2013*	IVF PGT-A cycles for: AMA (n = 102), RIF (n = 16), RM (n = 7)	138	455	Embryoscope, 15 min, *tPNf, t2, t3, t4, t5, t8, tSC, tSB, tB, cc2, cc3, S2, t5-t2, S3*	aCGH, Day 5/6 Trophectoderm (59%)	Not given ✓ *(logistic regression analysis demonstrated no correlation between aneuploidy and morphokinetics)*	Unknown ✗	✗	No difference in morphokinetic variables for aneuploid vs. euploid embryos. Tests the Basile model: not able to discriminate euploid embryos.	Yes
Patel *et al.*, 2016, India, Retrospective Cohort, *October 2013 to April 2015*	IVF PGT-A cycles for: AMA (n = 9), RIF (n = 1), RM (n = 7)	26	167	Embryoscope, Unknown, *tPB2, tPN, tPNf, t2, t3, t4, t5, t6, t7, t8, t9, tM, tSB, tB, tEB, tHB, cc2, cc3, S2, t5-t2*	aCGH, Day 3 blastomere, (75%)	32.9 (3.19) ✗	Unknown ✗	✗	No difference in morphokinetic variables for aneuploid vs. euploid embryos. Tested the Basile model: no significant difference.	Yes
Minasi *et al.*, 2016, Italy, Retrospective Cohort *September 2012 to April 2014*	IVF PGT-A cycles, indications unknown	454	928	Embryoscope, Unknown, *tPB2, tPN, tPNf, t2, t3, t4, t5, t6, t7, t8, t9, tM, tSB, tB, tEB, tHB, cc2, S2*	aCGH, Day 3 blastomere, (75%)	36.8 (4.2) ✗ *(only adjusted for morphological parameters)*	Unknown ✗	✗	Timing of cleavage from 3–4 cells, start of blastulation, expansion and hatching were significantly longer in aneuploid embryos.	Yes
Mumusoglu *et al.*, 2017, Turkey, Retrospective Cohort, *April 2015 to April 2016*	IVF PGT-A cycles for: AMA (n = 87), Genetic disorders (n = 16)	103	415	Embryoscope, 15 min, *tPB2, tPN, tPNf, t2, t3, t4, t5, t6, t7, t8, t9, tM, tSB, tB, tE, cc3, S2, t5-t2, S3, Blastulation*	aCGH, Day 5/6 Trophectoderm, (58%)	38 (4.7) ✓ *(age did not have a significant impact upon Morphokinetic variables)*	25.2 (4.3) ✓ *(BMI significantly affected tPNa, tPNF, t2, t4, t5, t6, t7, t8 and S, after adjustment only 5 parameters remained significantly delayed in aneuploid)*	✓ *(total FSH dose was significantly related to tB, tEB and blastulation- the higher the dose, the longer to reach the morphokinetic timings)*	After adjustment for confounders five variables remained significantly delayed in aneuploid embryos compared to euploid: t9, tM, tSB, tEB.	Yes

Continued

Table II Continued

Morphokinetics and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factors-Method, assessment period, morphokinetic variables**	**Outcome measurement-PGT-A platform, stage and type of biopsy (aneuploidy rate)**	**Confounding factors and attempted adjustment yes (✓) or no (✗)**			**Main findings**	**Included in quantitative analysis (yes or no)**
						Mean age (SD)	**BMI (SD)**	**Stimulation**		
Del Carmen Nogales *et al.*, 2017, Spain, Retrospective Cohort, *March 2013 to August 2014*	IVF PGT-A cycles for: RM (n = 47), RIF (n = 34), AMA (n = 31)	112	485	Embryoscope, 15 min, *T3, t5, cc2, t5-t2*	aCGH, Day 5/6 Trophectoderm, (61%)	35.5 ✗	Unknown ✗	✗	No significant difference in morphokinetic variables for embryos with monosomy or trisomy. Embryos with complex chromosomal abnormalities (>1), t3 (>34.7 h) and t5-t2 (>21 h), more likely to be normal.	No- separates morphokinetics for monosomy, complex, trisomy and euploid
Zhang *et al.*, 2017, China, Retrospective Cohort, *October 2014 to September 2015*	IVF PGT-A cycles for: RM, RIF, AMA (proportions unknown)	72	256	Embryoscope, 15 min, *tPNF, t2, t3, t4, t5, t8, t9, tSC, tM, tSB, tB, tEB, cc2, cc3, S2, t5-t2, S3*	aCGH, Day 5/6 Trophectoderm, (41%)	33.6 (4.2) ✗	Unknown ✗	✗	No significant difference in morphokinetic variables between aneuploid and euploid embryos. Tested Campbell and Basile model: not able to discriminate normal embryos.	Yes
Desai *et al.*, 2018, USA, Retrospective Cohort, *April 2012 to June 2017*	IVF PGT-A cycles, indications unknown	130	767	Embryoscope, 15 mins, *t2, t3, t4, t5, t8, tM, tSB, tB, tEB, cc2, S2, t5-t2,*	aCGH or NGS, Day 5/6 Trophectoderm, (58%)	36.3 (4.3) ✓ (after logistic regression analysis: still significant for tSB, tEB, tEB-tSB variables)	Unknown ✗	✗	tSB, tEB and tEB-tSB significantly prolonged in aneuploid vs. euploid embryos.	Yes
Lee *et al.*, 2019, Taiwan, Retrospective Cohort, *January to December 2017*	IVF PGT-A cycles, RIF, RM, unexplained, male factor (proportions unknown)	108	408	Embryoscope, Unknown, *tPNf, t2 t3, t4, t5, t8, tM, tSB, tB, cc2, cc3, S2, S3, Blastulation*	hrNGS, Day 5/6 Trophectoderm, (22.8%)	32.5 (4.1) ✗ *(only adjusted for KID5 score)*	21.6 (2.9) ✗	✗	Aneuploid blastocysts exhibited a significant delay in tB	Yes
Kimelman *et al.*, 2019, USA, Retrospective Cohort, *2015–2016*	IVF PGT-A cycles, indications unknown	152	754	Embryoscope, Unknown, *tPNf, t2, t3, t4, t5, t8*	NGS or SNP arrays, D5/6 Trophectoderm, (43.5%)	36.8 ✗	24 ✗	✗	t7 and t8 are independent prezictors of aneuploid. Blastocysts with high level mosaicism had significant delayed in t5, t8 and cc3.	Yes
Martin *et al.*, 2021, Spain, Retrospective Cohort, *April 2018 to April 2019*	IVF PGT-A cycles: AMA (n = 221), SMF (n = 37)	Unknown	1511	Embryoscope and Geri, 15 and 5 min, respectively *tSC, tSB, tB*	NGS, D5/6 Trophectoderm, (47%)	36.8 (5.4) ✓ *(age did not have a significant effect on morphokinetics)*	22.3 (3.5) ✓ *(BMI did not have a significant effect on morphokinetic)*	✓ *(Stimulation dose did not have a significant effect on morphokinetics)*	Aneuploid embryos display significantly delayed tSB, tSB and tB compared to euploid embryos. Embryos morphokinetics were not correlated to the degree of mosaicism.	Yes

AMA; advanced maternal age; PA: previous child affected by aneuploidy; RIF: recurrent implantation failure; RM: recurrent miscarriage.

Table III Characteristics of included studies for fragmentation and ploidy.

Fragmentation and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factor-Fragmentation assessment time point, *thresholds used for degree of fragmentation* (%)**	**Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)**	**Confounding factors & Attempted adjustment yes (✓) or no (✗)**			**Main Findings**	**Included in the quantitative analysis (Yes or No)**
						Mean age (SD)	**BMI (SD)**	**Stimulation**		
Magli *et al.*, 2001, Italy, Retrospective Cohort, 1996–2000	AMA (176 cycles), RIF (60 cycles), Abnormal Karyotype (47 cycles) TESE patients (31 cycles), Other (29 cycles)	239	1596	Assessed 112 h post insemination, *5–10, 11–20, 21–30, 31–40, >40*	FISH, D3 blastomere (66%)	Unknown ✗	Unknown ✗	✗	Strong association between percentage of fragmentation and chromosome abnormality.	No- raw numbers not given in the figure therefore could not be extracted
Ziebe *et al.*, 2003, Denmark, Retrospective analysis of RCT data March 2000 to June 2001	Randomly selected pool of donated embryos	143	103	Assessed at 26, 44, 50 and 68 h post insemination *0, <10, 11–20, 21–50*	FISH, D3 blastomere (45%)	25–37 (range) ✗	Unknown ✗	✗	Percentage of fragmentation was significantly associated with an increase in the rate of aneuploidy when assessed at 68 h but not when assessed at 48 h. Embryos with <20% fragmentation resulted in an increase of euploidy form 53% to 67% ($P = 0.012$).	Yes
Delimitreva *et al.*, 2005, Bulgaria, Retrospective Cohort, Unknown	Donated embryos (108 after successful and 103 after failed IVF)	181	169	Assessed 50–76 h post insemination *0, <20, <50*	FISH, D3 blastomere (15%)	33 ✗	Unknown ✗	✗	There was no difference in ploidy status for embryos up to 20% fragmentation, embryos with 20–50% fragmentation were more likely to be chromosomally abnormal.	Yes
Magli *et al.*, 2007, Italy Retrospective Cohort, September 1996 to February 2004	AMA, RIF, RM	662	5227	Assessed at 62 h post insemination, *0–10, 11–20, 21–30, 31–40*	FISH, Blastomeres biopsied from embryos with at least four cells (66% 'chromosome abnormalities')	Unknown ✗	Unknown ✗	✗	The incidence of aneuploidy in embryos with 4–6 cells on Day 3 was similar irrespective of the degree of fragmentation. There was a significant increase in the prevalence of aneuploidy for seven and eight cell embryos with higher percentage of fragmentation.	Yes
Moayeri *et al.*, 2008, USA, Retrospective Cohort, August 2001 to January 2005	AMA, RIF, RM, SMF	144	1081	Assessed embryo with at least four blastomeres *Veecks Grading System*	FISH, D3 blastomere (67%)	37.7 ✗ Split into 'young' (34.2) and 'AMA' groups (40.1)	Unknown ✗	✗	Regression analysis demonstrated that fragmentation was the best morphological predictor of ploidy status in both young and AMA groups (P -0.003, $P = 0.034$, respectively). Higher aneuploidy rates with higher degree of fragmentation.	No- Used grading system involving other morphological parameters, no data for solely fragmentation

Continued

Table III Continued

Fragmentation and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factor-Fragmentation assessment time point, *thresholds used for degree of fragmentation* (%)**	**Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)**	**Confounding factors & Attempted adjustment yes (✓) or no (✗)**			**Main Findings**	**Included in the quantitative analysis (Yes or No)**
						Mean age (SD)	**BMI (SD)**	**Stimulation**		
Chavez *et al.*, 2012, USA, Retrospective Cohort, Unknown	Donated embryos from successful IVF cycles	Unknown	45	Assessed at 30 h post insemination *'high' and 'low' fragmentation*	aCGH, D2 Blastomere (85%)	33.5 ✗	Unknown ✗	✗	Higher degree of fragmentation was associated with higher probability of aneuploidy (73%) vs. low degree (40%)	No- 'high' vs. 'low' only
Vera-Rodriguez *et al.*, 2015 USA, Retrospective Cohort, Unknown	Donated embryos (prognosis unknown)	19	57	Assessed using multi-plane imaging *0–5, 6–10, 11–15, 16–20, 21–25, 26–30, 31–35, 36–40, 41–45, 46–50*	aCGH, D3 blastomere, (51%)	33.7 (4.3) ✗	Unknown ✗	✗	Fragmentation not associated with ploidy status	No- When testing differences in ploidy compared <25% and >25% fragmentation only
Minasi *et al.*, 2016 Italy, Consecutive Case Series, September 2012 to April 2014	IVF PGT-A cycles, indications for PGT-A unknown	454	1730	Assessed Day 3 embryos *≤20, 20–50%, >50%*	aCGH, D3 blastomere, (65%)	36.8 (4.24) ✗	Unknown ✗	✗	Fragmentation not associated with ploidy status	Yes

AMA: advanced maternal age; PA: previous child affected by aneuploidy; RIF: recurrent implantation failure; RM: recurrent miscarriage.

Table IV Characteristics of included studies for abnormal cleavage and ploidy.

Abnormal Cleavage and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factor-Abnormal Cleavage Assessment Method and Cleavage patterns assessed**	**Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)**	**Confounding factors & Attempted adjustment, yes (✓) or no (✗)**			**Main Findings**	**Included in the quantitative analysis (Yes or No)**
						Mean age (SD)	**BMI (SD)**	**Stimulation**		
Campbell *et al.*, 2013a, UK, Retrospective Cohort May 2011 to July 2012	RIF, RM, SMF, Previous aneuploidy, AMA	25	98	Embryoscope, DUC1 and DUC2	aCGH or SNP, D5/6 Trophectoderm (61%)	38.6 (3.6) ✗	Unknown ✗	✓ (No dose dependent change in rate of abnormal cleavage)	No significant difference between aneuploid and euploid embryos for the presence of irregular division patterns	Yes
Rienzi *et al.*, 2013, Italy, Abstract- Retrospective Cohort, Unknown	Unknown indication for PGT-A	64	295	Embryoscope, Pooled abnormal cleavage, second cell division <4 h	aCGH, D3 or 5, unknown biopsy type (71%)	Unknown ✗	Unknown ✗	✗	No significant difference between aneuploid and euploid embryos for the presence of irregular division patterns	Yes
Davies *et al.*, 2014, Greece, Abstract- Retrospective Cohort 2011–2013	Unknown indication for PGT-A	94	456	Embryoscope, Pooled abnormal cleavage	Unknown, D3 or 5, unknown biopsy type (Unknown)	Unknown ✗	Unknown ✗	✗	Irregular division patterns were seen predominantly in abnormal embryos (57%) vs. in euploid embryos (19%) ($P = 0.01$)	No: no data to be able to calculate RR
Vera-Rodriguez *et al.*, 2015, USA, Retrospective Cohort, Unknown	Donated embryos, prognosis unknown	19	57	TS Auxogyn, pooled abnormal cleavage	aCGH, D3 blastomere, (51%)	33.7 (4.3) ✗	Unknown ✗	✗	The euploid rate was significantly lower in embryos displaying abnormal cleavage	Yes
Zhan *et al.*, 2016, USA, Retrospective Cohort, Unknown	Unknown indication for PGT-A	Unknown	1434	Embryoscope, DUC1-3	FISH/PCR/aCGH/SNP, D3 blastomere or D5 trophectoderm biopsy (38%)	Unknown ✓ (Occurrence of abnormal cleavage was not correlated with female age)	Unknown ✗	✗	The euploid rate was significantly lower in embryos displaying abnormal cleavage	Yes
Lagalla *et al.*, 2017, Italy, Retrospective Cohort, May 2013 to January 2015	SMF, AMA, RM, RIF	141	276	Embryoscope, DUC1, 2 and RV	aCGH, D5 trophectoderm (Unknown)	39.9 (4.4) ✓ (Occurrence of abnormal cleavage was not correlated with female age)	Unknown ✗	✗	Euploid embryos strongly associated with abnormal cleavage. Suggest that embryos that display AC but subsequently form blastocysts are more likely to be euploid. Significant number of embryos in this study arrested before biopsy (78%).	Yes

Continued

Table IV Continued

Abnormal Cleavage and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factor-Abnormal Cleavage Assessment Method and Cleavage patterns assessed**	**Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)**	**Confounding factors & Attempted adjustment, yes (✓) or no (✗)**			**Main Findings**	**Included in the quantitative analysis (Yes or No)**
						Mean age (SD)	**BMI (SD)**	**Stimulation**		
Zhang *et al* 2017, China, Retrospective Cohort, *October 2014 to September 2015*	IVF PGT-A cycles for: RM, RIF, AMA (proportions unknown)	72	256	Embryoscope, Pooled AC and RC	aCGH, Day 5/6 Trophectoderm, (41%)	33.6 (4.2) ✗	Unknown ✗	✗	No significant difference between aneuploid and euploid embryos for the presence of irregular division patterns	Yes
Desai *et al.*, 2018, USA, Retrospective Cohort April 2012 to June 2016	Unknown indication for PGT-A	130	767	Embryoscope, DUC, Irregular Chaotic cleavage and RV	aCGH, D5/6 Trophectoderm (58.4%)	36.3 (4.3) ✗	Unknown ✗	✗	No significant difference between aneuploid and euploid embryos for irregular division patterns except when two or more dysmorphisms (including multinucleation) coexist.	Yes
Ho *et al.*, 2018, USA, Retrospective Cohort, *Unknown*	Embryos donated for research, prognosis unknown	5	57	Miri TLS, pooled abnormal cleavage	NGS, Blastomere Day 3 and Trophectoderm Day 6 (30%)	39.1 (2.3) ✗	Unknown ✗	✗	The euploid rate was significantly lower in embryos displaying abnormal cleavage	Yes
McCoy *et al.*, 2018, USA, Retrospective Cohort Unknown	AMA, RM, Sex selection, RIF, SMF, Aneuploidy carrier	Unknown	77	Embryoscope, DUC1-3	SNP/FISH/PCR/ aCGH, D3 blastomere or D5 Trophectoderm	Median= 37 (range 18–48) ✗	Unknown ✗	✗	The euploid rate was significantly lower in embryos displaying abnormal cleavage	Yes
Ozbek *et al.*, 2021, Turkey, Retrospective Cohort April 2015 to October 2017	AMA (n = 497), Monogenetic disease (n = 57)	554	1015	Embryoscope DUC1-3, RC	aCGH/NHS, D5/6 Trophectoderm, (65%)	36.8 (4.7) ✓ (not correlated with female age)	Unknown ✗	✓ No difference.	No significant difference between aneuploid and euploid embryos for the presence of irregular division patterns	Yes

AMA: advanced maternal age; PA: previous child affected by aneuploidy; RIF: recurrent implantation failure; RM: recurrent miscarriage.

Table V Characteristics of included studies for contraction and ploidy.

					Contraction and Aneuploidy					
						Confounding factors & Attempted adjustment, yes (✓) or no (✗)				
Reference, year, country, study design, sample period	Sample population	Number of patients	Number of embryos (with PGT-A results)	Prognostic factor-Contraction Assessment Method	Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)	Mean age (SD)	BMI (SD)	Stimulation	Main Findings	Included in the Quantitative ananalysis
Vinals Gonzalez et al., 2018 UK, Retrospective Cohort, January 2016 to October 2017	IVF PGT-A cycles, indications unknown	190	869	Embryoscope	NGS, D5/6 Trophectoderm	38 (2.9) ✗	24.4 (4.4) and 23.1 (3.8) in contracting and non-contracting groups, respectively	✗	Aneuploid embryos displayed a significantly higher number of contractions compared to euploid embryos ($P < 0.001$)	Yes
Gazzo et al., 2020 Peru, Retrospective Cohort, Unknown	IVF PGT-A cycles, indications unknown	270	778	Embryoscope	NGS, Unknown stage, Trophectoderm	30.4 (range = 24–39) ✗ No correlation between age and contractions	Unknown ✗	✗	Aneuploid embryos displayed a significantly higher number of contractions compared to euploid embryos ($P = 0.029$)	Yes

Yilmaz et al., 2014). In contrast, three studies reported no association with multinucleation when examined during daily morphology assessments (Davies et al., 2012; Goodman et al., 2016; Del Carmen Nogales et al., 2017) or at the two-cell stage (Li et al., 2015). The findings of the main analysis were unchanged when excluding studies with the highest risk of bias (Supplementary Fig. S7).

Discussion

Key findings

Our study has found that aneuploid embryos are, on average, delayed by ≥1 h in t8 and ≥2 h in the morphokinetic variables t9 and tEB. Overall, in the weighted mean difference analysis, seven morphokinetic variables were significantly delayed in aneuploid embryos (tPB2, t2, t4, t7, t8, t9 and tEB). Blastocysts displaying contractions are associated with aneuploidy and reverse cleavages are more prevalent in euploid embryos, although these results should be interpreted with caution and investigated further before any conclusions can be drawn. In addition, although not statistically significant, there is a trend towards aneuploid embryos displaying multinucleation persisting to the four-cell stage. The trend between increasing percentage fragmentation and aneuploidy needs confirming in future studies owing to very low-quality evidence.

Morphokinetics and ploidy

Since the development of the Campbell model, there has been a plethora of attempts to test and create models for ploidy status, each with significant limitations (Campbell et al., 2013b). This original model has been tested by several authors; only Desai et al. (2018) was able to reliably risk stratify for aneuploidy (Kramer et al., 2014; Rienzi et al., 2015; Zhang et al., 2017; Desai et al., 2018). This may be because the morphokinetics of embryos are so sensitive to laboratory conditions that models may not be translatable between clinics or patient populations. Indeed, this variability may also account for why some models incorporate early cleavage parameters (Chavez et al., 2012; Chawla et al., 2015; Patel et al., 2016; Del Carmen Nogales et al., 2017) and some late, blastulation variables (Campbell et al., 2013a; Kramer et al., 2014; Desai et al., 2018; Lee et al., 2019; Martin et al., 2021). There are several common limitations to the published models, including the lack of control of confounders and the use of apparent validation by some authors, leading to model overestimation (Basile et al., 2014; Chawla et al., 2015; Del Carmen Nogales et al., 2017; Desai et al., 2018). In fact, confounding variables were overlooked in over 70% of the included studies (Supplementary Fig. S1). Four articles attempted to adjust for age, finding no association between age and morphokinetics (Rienzi et al., 2015; Desai et al., 2018; Martin et al., 2021; Mumusoglu et al., 2017). Conversely, BMI was found to be associated with delayed morphokinetics, while yet another such study demonstrated no such association (Martin et al., 2021; Mumusoglu et al., 2017). The effects of stimulation dosages were only assessed by three authors; two concluded there were no dose dependant differences, whereas one reported higher dosages were associated in delayed development kinetics (Campbell et al., 2013a; Martin et al., 2021; Mumusoglu et al., 2017).

Table VI Characteristics of included studies for multinucleation and ploidy.

Multinucleation and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factor-Multinucleation Assessment Method**	**Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)**	**Confounding factors & Attempted adjustment, yes (✓) or no (✗)**			**Main Findings**	**Included in the Quantitative Analysis (Yes or No)**
						Mean age (SD)	**Mean BMI (SD)**	**Stimulation**		
Kligman *et al.*, 1996 USA, Retrospective Cohort, Unknown	IVF PGT-A cycles, indications unknown	Unknown	450	Standard Morphology assessment D2, D3	FISH (XY, 18, 13, 21), D3 blastomere	35.3, 36.4 ✓ *(No significant difference in age for MN or non MN patients, respectively)*	Unknown ✗	✗	No significant difference between multinucleated and non-multinucleated embryos	Yes
Magli *et al.*, 2001, Italy, Retrospective Cohort, September 1996 to April 2020	AMA (176 cycles), RIF (60 cycles), Abnormal Karyotype(47 cycles) TESE patients (31 cycles), Other (29 cycles)	256	1489	Standard morphology assessment 40, 62, 88 and 112 h post insemination	FISH, D3 blastomere	Unknown ✗	Unknown ✗	✗	Multinucleation is significantly associated with complex chromosomal abnormalities.	Yes
Agerholm *et al.*, 2008 Denmark, Retrospective Cohort Unknown	Donated surplus embryos	35	35	Morphology analysis software, FertiMorph	FISH, D3 blastomere biopsy	Unknown ✗	Unknown ✗	✗	Presence of binucleated blastomeres in an embryo indicates increased risk of chromosome abnormalities	Yes
Scott *et al.*, 2010, USA, Abstract-Retrospective Cohort, Unknown	IVF PGT-A cycles, indications unknown	Unknown	12244	Standard morphology assessment every day	Unknown, D3 Unknown biopsy type	Unknown ✓ *(Age did not affect occurrence of MN)*	Unknown ✓ *(BMI had no affect on MN)*	✗ *(flare protocol used more in those patients with MN $p < 0.01$)*	Multinucleation associated with 4-fold increase in aneuploid rate on Day 3	No: data to calculate RR not given
Ambroggio *et al.*, 2011, USA, Retrospective Cohort, January 2004 to December 2009	RM or RIF	Unknown	141	Assessed Day 2 embryos using the Embryoscope	FISH (XY, 13, 15, 16, 18, 21, 22), D3 blastomere	38.2 (4), 38.9 (3.5) ✓ *(No significant difference in age for MN or non MN patients, respectively)*	Unknown ✗	✓ *(No difference in FSH and peak E2 level between MN and Non-MN)*	Significantly increased incidence of aneuploid in Day 2 multinucleated embryos when compared to non-multinucleated.	Yes
Davies *et al.*, 2012, Greece, Abstract- Retrospective Cohort, Unknown	IVF PGT-A cycles, indications unknown	14	70	Assessed cleavage stage embryos using the Embryoscope	aCGH & NGS, Cleavage stage single cell blastomere biopsy	Unknown ✗	Unknown ✗	✗	No significant difference in ploidy status for the presence of multinucleation	No: data to calculate RR not given
Campbell *et al.*, 2013a, UK, Retrospective Cohort, May 2011 to May 2012	RIF, RM, SMF, Previous aneuploidy, AMA	25	98	Assessed at 2 cell and 4 cell stage using the Embryoscope	aCGH or SNP array, 70% D3, 30% D5/6, Trophectoderm biopsy	36.6 (5.1) ✗	Unknown ✗	✓ *(No dose dependent differences for ploidy results)*	No significant difference in ploidy status for the presence of multinucleation at the 2 cell stage	Yes

Continued

Table VI Continued

Multinucleation and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factor-Multinucleation Assessment Method**	**Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)**	**Confounding factors & Attempted adjustment, yes (✓) or no (✗)**			**Main Findings**	**Included in the Quantitative Analysis (Yes or No)**
						Mean age (SD)	**Mean BMI (SD)**	**Stimulation**		
Mazur *et al.*, 2013, Ukraine, Abstract-Retrospective Cohort, March 2012 to January 2013	IVF PGT-A cycles, RIF, RM, abnormal karyotype	43	223	Assessed at 2 cell and 4 cell stage using the Embryoscope	aCGH, D5/6 Trophectoderm biopsy	32 ✗	Unknown ✗	✗	Multinucleation was equally present in euploid and aneuploid embryos but the frequency of severe multinucleation was twice higher in the aneuploid group.	Yes
Melzer *et al.*, 2013, USA, Abstract-Prospective Cohort, Unknown	IVF PGT-A cycles, indications unknown	20	213	Assessed at the 4 cell stage using the Embryoscope	aCGH, D5/6 Trophectoderm biopsy	Unknown ✗	Unknown ✗	✗	More aneuploid embryos were multinucleated at the 4 cell stage compared to euploid.	No: data to calculate RR not given
Munoz *et al.*, 2014, Spain, Abstract-Retrospective Cohort, September 2011 to January 2014	IVF PGT-A cycles, indications unknown	96	564	Assessed at 2 cell and 4 cell stage using the (Unknown) TLS	aCGH, unknown 'embryos of at least 6 cells biopsied'	Unknown ✗	Unknown ✗	✗	Multinucleation at the two cell stage is a transitory event. Multinucleation at the four cell stage is significantly associated with aneuploidy.	Yes
Yilmaz *et al.*, 2014, Canada, Retrospective Cohort, March 1998 to November 2011	IVF PGT-A cycles, indications unknown	126	189	Microscopy and spreading and staining of nuclei	FISH, D3 blastomere biopsy	31.6 (2.6) ✗	Unknown ✗	✗	Multinucleation is significantly associated with aneuploid.	No: data to calculate RR not given
Bayram *et al.*, 2015, UAE, Abstract- Retrospective Cohort, January 2014 to December 2014	IVF PGT-A cycles, indications unknown	50	261	Assessed at 2 cell and 4 cell stage using the Embryoscope	aCGH, D3 unknown cell type	34.9 ✗ *Similar mean ages across groups*	Unknown ✗	✗	Multinucleation at the two and four cell stage significantly associated with aneuploidy	Yes
Li *et al.*, 2015, USA, Abstract- Retrospective Cohort, Unknown	IVF PGT-A cycles, indications unknown	86	24	Assessed at 2 cell stage using the Embryoscope	aCGH, blastomere biopsy (n = 152) or trophectoderm (n = 86)	(range only 26–44) ✗	Unknown ✗	✗	Multinucleation at the two-cell stage not associated with increased occurrence of aneuploidy	No: data to calculate RR not given
Balakier *et al.*, 2016, Canada, Retrospective Cohort Unknown	AMA (n = 49), PCOS (n = 18), SMF (n = 39), Some embryos from donated oocytes (n = 98)	113	607	Assessed at 2 cell and 4 cell stage using the Embryoscope	aCGH, D5/6 Trophectoderm biopsy	37.62 ✗ (5.18) for non-donors, 25.7 (3.6) donors	Unknown ✗	✓ *(no difference in E2 level on day of trigger)*	No significant difference in ploidy status for the presence of multinucleation	Yes

Continued

Table VI Continued

Multinucleation and Aneuploidy										
Reference, year, country, study design, *sample period*	**Sample population**	**Number of patients**	**Number of embryos (with PGT-A results)**	**Prognostic factor-Multinucleation Assessment Method**	**Outcome measurement-PGT-A platform, stage and type of biopsy (Aneuploidy rate)**	**Confounding factors & Attempted adjustment, yes (✓) or no (✗)**			**Main Findings**	**Included in the Quantitative Analysis (Yes or No)**
						Mean age (SD)	**Mean BMI (SD)**	**Stimulation**		
Zhang *et al.*, 2017, China, Retrospective Cohort, *October 2014 to September 2015*	IVF PGT-A cycles for: RM, RIF, AMA (proportions unknown)	72	256	Assessed cleavage stage embryos using the Embryoscope	aCGH, Day 5/6 Trophectoderm, (41%)	33.6 (4.2) ✗	Unknown ✗	✗	Multinucleation is significantly associated with aneuploid.	Yes
Goodman *et al.*, 2016, USA, Abstract- Prospective Cohort, Unknown	Unexplained infertility patients with PGT-A cycles	9	133	Assessed on Day 2 using the Embryoscope	Unknown, D5/6 Trophectoderm biopsy	31.9 (3.2) ✗	Unknown ✗	✗	No significant difference in ploidy status for the presence of multinucleation when assessed on Day 2	No: data to calculate RR not given
Hashimoto *et al.*, 2016, Japan, Retrospective Cohort, Unknown	Donated embryos, prognosis unknown	44	26	Confocal imaging studies with RNA Chromosome tracking assessed at 4 cell stage	aCGH, Blastocyst, cell type unknown	34.8 (4.3) ✗	Unknown ✗	✗	No significant difference in ploidy status for the presence of multinucleation when assessed at the four-cell stage	Yes
Del Carmen Nogales *et al.*, 2017, Spain, Abstract- Retrospective Cohort Unknown	RM or RIF	Unknown	478	Assessed on Day 2 using the Embryoscope	aCGH, D3 Unknown cell type	Unknown ✗	Unknown ✗	✗	No significant difference in ploidy status for the presence of multinucleation.	No: data to calculate RR not given
Desai *et al.*, 2018, USA, Retrospective Cohort, April 2012 to June 2016	IVF PGT-A cycles, indications unknown	130	767	Assessed cleavage stage embryos using the Embryoscope	aCGH, D5/6 Trophectoderm	36.3 (4.3) ✓ Adjusts for female age in Logistic Regression analysis	Unknown ✗	✗	No significant difference in ploidy status for the presence of multinucleation.	Yes
Lee *et al.*, 2019, Taiwan, Retrospective Cohort, *January to December 2017*	VF PGT-A cycles, RIF, RM, unexplained, male factor (proportions unknown)	108	408	Assessed cleavage stage embryos using the Embryoscope	hrNGS, Day 5/6 Trophectoderm, (22.8%)	32.5 (4.1) ✗ *(only adjusted for KID5 score)*	21.6 (2.9) ✗	✗	No significant difference in ploidy status for the presence of multinucleation.	Yes

AMA: advanced maternal age; PA: previous child affected by aneuploid; RIF: recurrent implantation failure; RM: recurrent miscarriage.

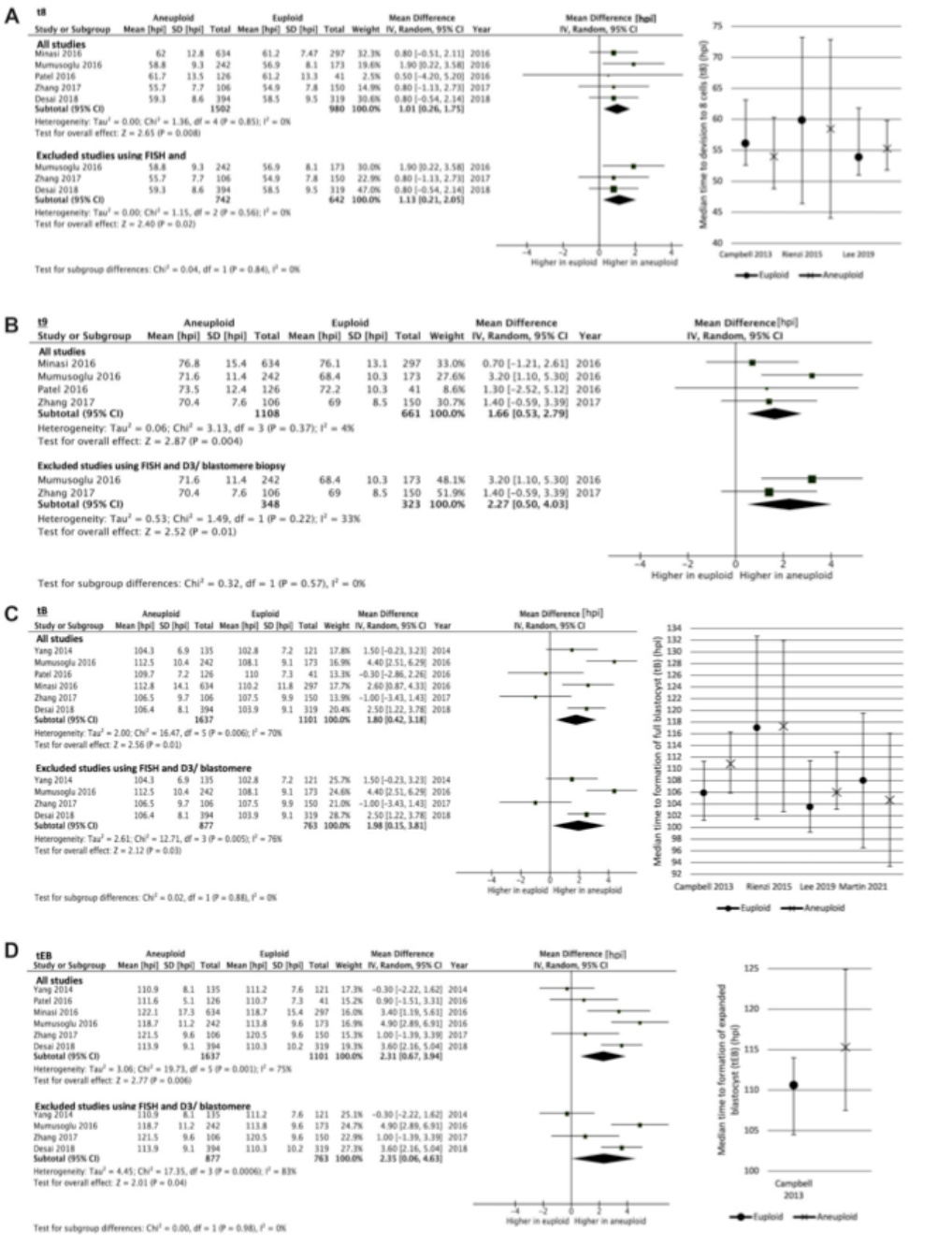


Figure 6. Weighted mean difference and prognostic factor analysis graphs of aneuploid versus euploid human embryos for morphokinetic variables. hpi, hours post-insemination; t8 **(A):** time from insemination to 8 cells (hpi); t9 (**B):** time from insemination to 9 cells (hpi); tB **(C):** time from insemination to the formation of a full blastocyst (hpi); tEB (**D**): time from insemination to expanded blastocyst (hpi).

Figure 7. Relative risk of euploidy in a human embryo displaying reverse cleavage. M-H, Mantel–Haenszel analysis; RC, reverse cleavage.

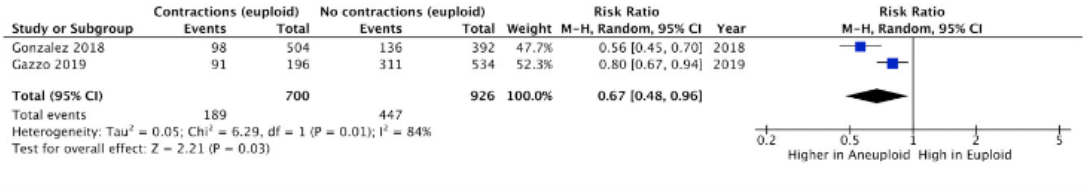


Figure 8. Relative risk of euploidy in a human embryo displaying contractions. M–H, Mantel–Haenszel analysis.

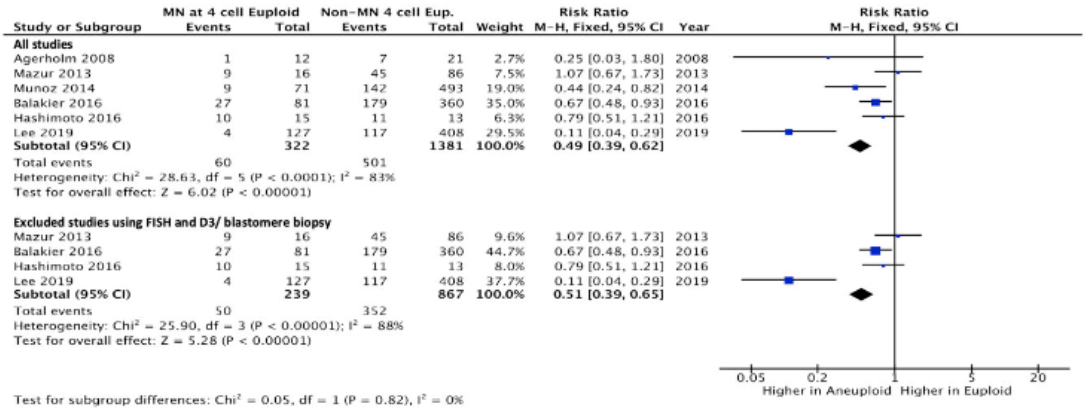


Figure 9. Relative risk of a multinucleated human embryo being euploid when assessed at the four-cell stage. M–H: Mantel–Haenszel analysis; MN: multinucleated.

In comparison to t8, t9 and tEB, the variables tPB2, t2, t4 and t7 were less dramatically delayed in aneuploid embryos; therefore, in the context of such wide CIs are less likely to reliably predict ploidy status. Of the most delayed variables, t8 and t9 had minimal heterogeneity ($I^2 = 0\%$ and 33%, respectively), whereas tB and tEB were substantially heterogenous ($I^2 = 76\%$ and 83%, respectively). The reasons for the

heterogeneity are multifactorial including diverse patient populations, insufficient control for confounders, lack of standardization of morphokinetic annotations, differences in laboratory and genetic testing techniques, and diverse embryo culture conditions. It must be highlighted that the results from tB and tEB are significantly heterogenous, therefore conclusions regarding these variables cannot reliably be drawn. That said, the heterogeneity for tEB is trending towards aneuploidy rather than traversing across the line of no effect. Ordinarily, we would be opposed to the meta-analysis of such heterogenous results; however, the aim of this systematic review was not to provide a summary statistic to be translated directly into model development but to indicate potential prognostic markers for testing at local units. Whilst they are heterogenous, the results highlight the trend towards blastulation parameters predicting aneuploidy but, that said, we acknowledge that further research is needed to confirm our findings for tB and tEB.

The sensitivity analysis did not alter the results when studies with a high risk of bias were excluded; however, tSB became significant with the use of a fixed effects model. This would indicate the need for more data to reliably conclude whether this variable could act as a prognostic marker. Interestingly, two morphokinetic studies were of higher quality and had comparable findings to our conclusions (Martin *et al.*, 2021; Mumusoglu *et al.*, 2017).

The association of day of blastocyst formation with aneuploid rates has been extensively studied, illustrating an increasing prevalence of aneuploidy from Day 5 to Day 7 blastocysts (Whitney *et al.*, 2016; Minasi *et al.*, 2016; Su *et al.*, 2016; Werland *et al.*, 2017; Kaing *et al.*, 2018; Hernandez-Nieto *et al.*, 2019; Tiegs *et al.*, 2019; McDaniel *et al.*, 2021). Critics of time-lapse technology would argue that there is little to be gained from the study of cleavage parameters over the day of blastocyst formation using traditional monitoring. We argue that whilst day of blastocyst formation is a useful tool to counsel patients with limited access to time lapse, the accuracy and practicality that a TLS offers (for assessing readiness for biopsy whilst remaining in culture) is irreplaceable. Relying solely on traditional methods can lead to inaccuracies with the timing of blastulation in comparison to a TLS, where t0 is standardized to tPNf or time post insemination (hpi), allowing a more precise discrimination of a viable embryo despite slower development. The most successful morphokinetic logistic regression models for live birth are now much more complex than those using predefined thresholds, such as tSB <116 h or more traditional hierarchical models (Petersen *et al.*, 2016; Zaninovic *et al.*, 2017; Fishel *et al.*, 2018). Time lapse therefore allows a statistical interpretation of embryo development whilst accounting for confounders that is not possible using traditional methods or univariate analysis. The variables more confidently associated with aneuploidy in this review are t8 and t9, factors that can only be considered through time lapse. It must also be considered that whilst tEB showed prognostic potential for ploidy, tSB and tB were not significantly associated, highlighting the precise nature of these associations rather than simply blastocyst formation. Finally, it has been suggested that there is some degree of multi-collinearity between cleavage and blastocyst kinetics, and this is illustrated by the fact that several authors have used earlier variables to predict blastocyst development (Wong *et al.*, 2010; Cruz *et al.*, 2012; Dal Canto *et al.*, 2012; Hashimoto *et al.*, 2012; Kirkegaard *et al.*, 2013a; Desai *et al.*, 2014; Milewski *et al.*, 2016). Therefore, this raises the question as to whether cleavage variables add prognostic value over the later blastulation parameters. Unfortunately, this has not been directly compared as published models either incorporate early cleavage parameters (Chavez *et al.*, 2012; Chawla *et al.*, 2015; Patel *et al.*, 2016; Del Carmen Nogales *et al.*, 2017) or blastulation variables (Campbell *et al.*, 2013a; Kramer *et al.*, 2014; Desai *et al.*, 2018; Lee *et al.*, 2019; Martin *et al.*, 2021). This would be an interesting question to drive future research, and care would need to be taken to not 'cherry-pick' variables to be included in prognostic model development, however, as this can introduce significant bias outside the context of prognostic factor research (Riley *et al.*, 2019).

More recently, artificial neural networks have demonstrated an impressive ability to evaluate images of pre-implantation embryos. Chavez-Badiola *et al.* (2020) developed a ranking system for ploidy status using this technology, with an impressive AUC of 0.70. Interestingly, two groups have investigated if there was an additive effect of using morphokinetic algorithms with artificial intelligence to improve diagnostic accuracy (Barnes *et al.*, 2020; Huang *et al.*, 2021). Barnes *et al.* (2020) demonstrated that both work synergistically to improve the AUC from 0.62 when solely image analysis is used to 0.76 (Barnes *et al.*, 2020). Huang *et al.* (2021) similarly found the AUC increased from 0.57 to 0.77 with the addition of morphokinetics, age and full video analysis. This use of artificial intelligence in combination with morphokinetic models is a new direction of research that is evolving. Initial results appear promising and further studies are needed to demonstrate the application of this methodology. It would be beneficial for future work to include a prospective study design to validate these more complex models.

Morphological features and ploidy

It has been established that embryos with higher degrees of fragmentation have lower implantation rates; if the relationship suggested by our results is in fact true, the aetiology may be, in part, due to aneuploidy (Ziebe *et al.*, 1997; Ebner *et al.*, 2001). The quality of the evidence presented in all studies is poor, predominantly because of the use of unreliable genetic technologies (all used blastomere biopsy of intact cells and many adopted the use of FISH). Furthermore, the characteristics of the included patients are also extremely heterogenous. Some studies include couples with a good prognosis, in contrast to others focussing on patients with recurrent miscarriage or advanced maternal age, with no methods used to account for this (Tables II–VI). Notably, the fragmentation assessment method and timing were also inconsistent. This is important given that one author concluded that fragmentation was only associated with aneuploidy when assessed at the seven and eight cell stages and others when assessed at 48 h (Ziebe *et al.*, 2003; Magli *et al.*, 2007). Other authors categorized fragmentation as 'high' or 'low'; these arbitrary thresholds make testing association more unreliable and to our knowledge, there is no evidence to support such an approach (Chavez *et al.*, 2012; Vera-Rodriguez *et al.*, 2015). All considered, we cannot reliably conclude whether percentage fragmentation is associated with aneuploidy. There is, therefore, a need for future adequately powered studies to examine fragmentation using time-lapse, next generation sequencing and with adequate control of confounding.

Reverse cleavage has been associated with euploidy in our results, but it should be considered that these findings come from the contribution of one study and all other authors concluded that there was no significant difference (Ozbek *et al.*, 2021). Whilst this was the largest

study with 8% of embryos (n = 78/1015) displaying reverse cleavage, the event rate remains low. For instance, we have calculated that for a power of 80% and a value of 0.05 for alpha, you would need a sample size of 1617 embryos with at least 147 displaying reverse cleavage in order to find a difference when one truly exists. This is presuming a difference of 12% in the euploid rate between embryos displaying reverse cleavage and those that did not (extrapolated from the studies in this meta-analysis) and assuming a 1:10 ratio for the presence of this dysmorphism to normal cleavage. This illustrates a significant limitation of studies investigating dysmorphisms with such low prevalence. Ozbek *et al.* (2021) provide no explanation why embryos displaying reverse cleavage may have a higher incidence of euploidy, particularly in the context of the dramatically inferior live birth rates stated in their study when compared with normally cleaved euploid embryos (23% versus 56%). This association between reverse cleavage and inferior implantation rates has been replicated by several other authors; therefore, we highly doubt that a relationship between euploidy and reverse cleavage truly exists (Liu *et al.*, 2014; Barrie *et al.*, 2017; Desai *et al.*, 2018), particularly considering the underpowered nature of this study and the fact that reverse cleavage is often associated with compromized embryo development and quality. In fact, studies of bovine embryos have demonstrated an association with aneuploidy, strengthening the argument that these results are likely spurious (Magata *et al.*, 2019). There have also been multiple factors independently associated with reverse cleavage, such as antagonist cycles, low progressive sperm motility and the use of ICSI (Liu *et al.*, 2014).

While our main analyses indicate that direct uneven cleavage is not associated with ploidy, there is a significant limitation to the designs of the included studies. Aneuploid embryos may have been inadvertently excluded, either because only good quality embryos were biopsied or because a significant proportion (up to 87%) arrested in their development before biopsy (Zhan *et al.*, 2016; Lagalla *et al.*, 2017). It would be safer to conclude that embryos that have displayed direct cleavage that make it to the blastocyst stage could still be considered for biopsy or transfer: it has been demonstrated that they can result in live births; however, the patient must be warned of the increased likelihood of adverse outcome (Fan *et al.*, 2016; Zhan *et al.*, 2016; Ozbek *et al.*, 2021). What causes these abnormal cleavages remains largely unknown, although it has previously been associated with the follicular environment of oocytes, poor-motility sperm and GnRH antagonists (Liu *et al.*, 2014). Considering this, and the fixed effects sensitivity analysis that demonstrated DUC2 to be significantly associated with aneuploidy, further investigation is required to confirm or refute these findings.

Embryo contraction is a common phenomenon observed in a TLS (42% of embryos in the included studies), yet despite an understanding of the physiology, causality remains controversial. It has been hypothesized that contractions may assist in embryo hatching, although recent evidence does not support this theory (Gazzo *et al.*, 2020). Future research should exclude studies that have undergone assisted hatching on Day 3 as this has been related to altered frequency of contractions, a limitation of the included studies in the current analysis (Vinals Gonzalez *et al.*, 2018; Gazzo *et al.*, 2020). Embryos displaying contractions were more likely to be aneuploid; however, these data comes from only two studies, therefore further research is recommended to investigate this association.

The relationship between multinucleation at the four-cell stage and ploidy is yet to be established given the significantly heterogenous results ($I^2 = 88\%$) and contradictory findings in the subgroup and sensitivity analysis. It has been described how the presence of multinucleation and associated aneuploidy can 'self-correct' by exclusion of cells during compaction or blastulation (Kligman *et al.*, 1996; Ambroggio *et al.*, 2011; Balakier *et al.*, 2016; Desai *et al.*, 2018). This complicates our understanding but may explain why only embryos displaying multinucleation at the four-cell stage may be associated with aneuploidy and how healthy babies have been born from such embryos (Meriano *et al.*, 2004; Yilmaz *et al.*, 2014). Furthermore, multinucleation is only visible at interphase during conventional culture, therefore is likely to be underreported in the five included studies not utilizing a TLS (Kligman *et al.*, 1996; Magli *et al.*, 2001; Agerholm *et al.*, 2008; Scott *et al.*, 2010; Ambroggio *et al.*, 2011). In addition to aneuploidy, the presence of multinucleation has been related to the use of agonist down-regulation (perhaps associating it with poor ovarian reserve), high FSH dosages, high oestrogen levels and excessive oocyte numbers (De Cássia Savio Figueira *et al.*, 2010; Scott *et al.*, 2010; Desai *et al.*, 2018). Despite this, across all the morphological studies, only two manuscripts reporting the use of statistical modelling to adjust for age and no other confounders were considered (Minasi *et al.*, 2016; Desai *et al.*, 2018). In contrast to embryos displaying abnormal cleavage, there has been no difference demonstrated in the development of multinucleated embryos to expanded blastocyst, therefore our results are unlikely to be affected by arrested embryos (Goodman *et al.*, 2016).

Strengths and limitations of this systematic review and meta-analysis

The findings of our study should be interpreted with caution due to an overall moderate risk of bias and significant heterogeneity of the included studies. Attempts have been made to control for sources of heterogeneity in our study design. This was primarily through subgroup analysis by excluding studies using older, unreliable technologies. In some variables, the heterogeneity was calculated to be worse in the subgroup than in the main analysis, and this highlights the manifestation of other factors contributing to the diversified results. Heterogeneity may also exist in the way studies classified mosaics; this definition remains ambiguous in several studies. This is important as mosaic embryos have previously been shown to have independent morphokinetic characteristics (Martin *et al.*, 2021). It is also worth considering that whilst it is generally accepted that PGT-A biopsy results are concordant with the rest of the embryo in most cases, it is not absolute and sceptics exist (Esfandiari *et al.*, 2016; Gleicher and Orvieto, 2017; Victor *et al.*, 2019). There have been reports and suggested models of so called 'self-correction mechanisms' whereby mosaic embryos become more chromosomally normal as development progresses, although the existence of this phenomenon remains debatable (Bolton *et al.*, 2016; Capalbo and Rienzi, 2017; McCoy, 2017; Munné *et al.*, 2017; Coticchio *et al.*, 2021a).

Of the studies included in the meta-analysis of morphokinetic variables, all used ICSI, thus timing development from insemination apart from two groups, namely Lee *et al.* (2019) and Chavez *et al.* (2012). Unfortunately, t0 remains ambiguous in the study by Chavez *et al.* (2012) due to unanswered correspondence. Lee *et al.* (2019) used

both standard IVF and ICSI for the included embryos, therefore this is a significant confounding factor to consider as they time conventional IVF embryos from the addition of spermatozoan to the oocyte; the accepted standard would be from tPNf. The exclusion of Chavez et al. (2012) for cc2 would make this variable not associated with ploidy status rather than associated with euploidy. That said, this study is not included in the subgroup analysis, therefore the findings for this variable and S2 remain unchanged when considering the studies using the most reliable genetic technology.

While the conclusions drawn from this study are taken from data of over 40 000 embryos, the quality of evidence is low due to imprecision and large CIs. Only a limited number of studies tested each variable, leading to low event rates for some variables and the inclusion of only a handful of studies of those reporting usable data. As discussed previously, this is even more profound when the sample size of patients rather than embryos is considered. Furthermore, if the true population mean lies on the lower boundary of the 95% CI, we would be unable to predict ploidy using tB and tEB. A final limitation is our inability to test for publication bias. That said, embryological studies tend to report a whole array of potential prognostic factors for ploidy per manuscript, therefore there is less risk of reporting only positive findings.

Our review does have multiple strengths, first the thorough methodological approach and comprehensive search of multiple variables and their association with ploidy status will be the first of its kind. Second, our meta-analyses of morphological and morphokinetic variables provide a strong argument for the local development of morphokinetic algorithms for ploidy and suggest those most likely to be included. Finally, we have provided an extensive critique of existing research and the quality of evidence in order to inform future prognostic methodologies.

Conclusion

In this first systematic review and meta-analysis of morphological and morphokinetic associations with ploidy, we have reported the most reliable prognostic markers to be t8, t9 and tEB. These results support the mechanism by which algorithms for live birth have predictive ability, suggesting that aneuploidy causes delayed cytokinesis. That said, we have demonstrated considerable variability within aneuploid and euploid embryos making definitively classifying them impossible. Time-lapse is, therefore, not suitable as a method to diagnose the ploidy status of pre-implantation embryos. Considering recent reports, it may be that morphokinetic algorithms can be used as a tool to risk stratify embryos for ploidy status, and more accurately by instrumenting artificial intelligence. Further research is needed to determine the suitability of machine learning for embryo assessment and selection.

Owing to the limited number of studies, heterogenous results and poor-quality evidence that suggested association between aneuploidy and multinucleation at the four-cell stage, frequency of embryo contractions and fragmentation needs to be investigated further. Adequately powered studies should be conducted to test our hypothesis that reverse cleavage is not associated with euploidy. We propose that incorporating associated morphological factors into a prognostic model may work synergistically to improve euploid embryo selection. On the other hand, multinucleation assessed on Day 2 or at the two-cell stage and direct unequal cleavage in an embryo that progresses to a blastocyst do not appear to be associated with ploidy.

Differing clinical and laboratory practices and inadequate control for confounders in previous research is most probably why TLS is rated as 'amber' by the UK regulatory body (HFEA, 2021b). There have been calls for multi-centre randomized controlled trials heard for many years (Armstrong et al., 2019). Instead, we argue that since embryos are so significantly affected by local conditions, it may be more appropriate to robustly test models developed in-house.

While this review concludes that a TLS cannot be used to definitively diagnose ploidy status, further research is needed to comprehend the potential of morphokinetic algorithms to prioritize embryos for biopsy, or to use morphokinetics to select between euploid embryos. Therefore, we will test this hypothesis in a cohort study at CARE Fertility using a morphokinetic dataset of over 8000 embryos with known PGT-A outcomes. This model will be trained, tested and validated geographically and, if successful, a prospective study will determine its discriminative ability. If successful, this has the potential to be a meaningful improvement for patients, aiming to make more advanced and costly reproductive technologies more accessible.

Supplementary data

Supplementary data are available at *Human Reproduction Update* online.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

Acknowledgements

The first author would like to thank CARE Fertility for providing a platform for research and clinical training through the creation of a clinical research fellow post. In particular, Mr Phillip Lowe (Medical Director, CARE Manchester), Dr Victoria Sephton (CARE Group Deputy Medical Director) and Professor Kingsland (CARE Group Clinical Director) for their continued support. The investigators would like to thank the following authors who supplied additional information about their research: Professor David McCulloh, Professor Magli, Dr Hakan Yarali, Dr Pavlo Mazur. Additionally, we would like to thank Dr Adam Devall for assistance in obtaining certain manuscripts.

Authors' roles

The study was conceived by Prof. Arri Coomarasamy and A.B. as part of a PhD programme of research undertaken by T.B. at the University of Birmingham. The study protocol was designed by T.B., S.M., and supervized by Prof. Arri Coomarasamy. Study selection and extraction of data was performed by T.B. and A.B., followed by a quality assessment by T.B. All authors analysed and interpreted the data. T.B. drafted the first manuscript, this was subsequently approved by all authors before publication. C.E. provided statistical support and A.C. provided expertize within this field of interest.

Funding

T.B. received funding obtained by CARE Fertility for PhD tuition fees and stipend at the University of Birmingham.

Conflict of interest

A.C. is a minor shareholder at CARE Fertility. No other conflicts of interest exist. To note, CAREmaps® is a technology for which patients are charged extra for at CARE.

References

Agerholm IE, Hnida C, Cruger DG, Berg C, Bruun-Petersen G, Kolvraa S, Ziebe S. Nuclei size in relation to nuclear status and aneuploidy rate for 13 chromosomes in donated four cells embryos. *J Assist Reprod Genet* 2008;**25**:95–102.

Alfarawati S, Fragouli E, Colls P, Stevens J, Gutiérrez-Mateo C, Schoolcraft WB, Katz-Jaffe MG, Wells D. The relationship between blastocyst morphology, chromosomal abnormality, and embryo gender. *Fertil Steril* 2011;**95**:520–524.

Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Hum Reprod* 2011;**26**:1270–1283.

Ambroggio J, Gindoff PR, Dayal MB, Khaldi R, Peak D, Frankfurter D, Dubey AK. Multinucleation of a sibling blastomere on day 2 suggests unsuitability for embryo transfer in IVF-preimplantation genetic screening cycles. *Fertil Steril* 2011;**96**:856–859.

Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database Syst Rev* 2019;**5**: CD011320.

Ata B, Kaplan B, Danzer H, Glassner M, Opsahl M, Tan SL, Munné S. Array CGH analysis shows that aneuploidy is not related to the number of embryos generated. *Reprod Biomed Online* 2012;**24**: 614–620.

Athayde Wirka K, Chen AA, Conaghan J, Ivani K, Gvakharia M, Behr B, Suraj V, Tan L, Shen S. Atypical embryo phenotypes identified by time-lapse microscopy: high prevalence and association with embryo development. *Fertil Steril* 2014;**101**:1637–1648.e1–5.

Balakier H, Sojecki A, Motamedi G, Librach C. Impact of multinucleated blastomeres on embryo developmental competence, morphokinetics, and aneuploidy. *Fertil Steril* 2016;**106**:608–614.e2.

Barnes J, Malmsten J, Zhan Q, Hajirasouliha I, Elemento O, Sierra J, Zaninovic N, Rosenwaks Z. Noninvasive detection of blastocyst ploidy (euploid vs. aneuploid) using artificial intelligene (AI) with deep learning methods. *Fertil Steril* 2020;**114**:e76.

Barrie A, Homburg R, McDowell G, Brown J, Kingsland C, Troup S. Preliminary investigation of the prevalence and implantation potential of abnormal embryonic phenotypes assessed using time-lapse imaging. *Reprod Biomed Online* 2017;**34**:455–462.

Barrie A, McDowell G, Troup S. An investigation into the effect of potential confounding patient and treatment parameters on human embryo morphokinetics. *Fertil Steril* 2021;**115**:1014–1022.

Basile N, Nogales MdC, Bronet F, Florensa M, Riqueiros M, Rodrigo L, García-Velasco J, Meseguer M. Increasing the probability of selecting chromosomally normal embryos by time-lapse morphokinetics analysis. *Fertil Steril* 2014;**101**:699–704.

Bayram A, Fakih M, Shunnar A, Divakaran J, Poukaram N. Does disappearance of multinucleation from the 2-cell to 4-cell stage observed by time time-lapse correlate chromosomal normality detected by PGS microarray? *Hum Reprod* 2015;**30**.

Bellver J, Mifsud A, Grau N, Privitera L, Meseguer M. Similar morphokinetic patterns in embryos derived from obese and normoweight infertile women: a time-lapse study. *Hum Reprod* 2013;**28**: 794–800.

Bodri D, Sugimoto T, Serna JY, Kawachiya S, Kato R, Matsumoto T. Blastocyst collapse is not an independent predictor of reduced live birth: a time-lapse study. *Fertil Steril* 2016;**105**:1476–1483. e1473.

Bolton H, Graham SJL, Van der Aa N, Kumar P, Theunis K, Fernandez Gallardo E, Voet T, Zernicka-Goetz M. Mouse model of chromosome mosaicism reveals lineage-specific depletion of aneuploid cells and normal developmental potential. *Nat Commun* 2016;**7**:11165.

Bronet F, Nogales MC, Martínez E, Ariza M, Rubio C, García-Velasco JA, Meseguer M. Is there a relationship between time-lapse parameters and embryo sex? *Fertil Steril* 2015;**103**:396–401.e2.

Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Hickman CF. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. *Reprod Biomed Online* 2013a;**26**: 477–485.

Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Thornton S. Retrospective analysis of outcomes after IVF using an aneuploidy risk model derived from time-lapse imaging without PGS. *Reprod Biomed Online* 2013b;**27**:140–146.

Capalbo A, Rienzi L. Mosaicism between trophectoderm and inner cell mass. *Fertil Steril* 2017;**107**:1098–1106.

Capalbo A, Rienzi L, Cimadomo D, Maggiulli R, Elliott T, Wright G, Nagy ZP, Ubaldi FM. Correlation between standard blastocyst morphology, euploidy and implantation: an observational study in two centers involving 956 screened blastocysts. *Hum Reprod* 2014; **29**:1173–1181.

Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reprod Biomed Online* 2020;**41**:585–593.

Chavez SL, Loewke KE, Han J, Moussavi F, Colls P, Munne S, Behr B, Reijo Pera RA. Dynamic blastomere behaviour reflects human embryo ploidy by the four-cell stage. *Nat Commun* 2012;**3**:1251.

Chawla M, Fakih M, Shunnar A, Bayram A, Hellani A, Perumal V, Divakaran J, Budak E. Morphokinetic analysis of cleavage stage embryos and its relationship to aneuploidy in a retrospective time-lapse imaging study. *J Assist Reprod Genet* 2015;**32**:69–75.

Ciray HN, Aksoy T, Goktas C, Ozturk B, Bahceci M. Time-lapse evaluation of human embryo development in single versus sequential culture media—a sibling oocyte study. *J Assist Reprod Genet* 2012;**29**:891–900.

Cohen J, Wells D, Munné S. Removal of 2 cells from cleavage stage embryos is likely to reduce the efficacy of chromosomal tests that are used to enhance implantation rates. *Fertil Steril* 2007;**87**: 496–503.

Coticchio G, Barrie A, Lagalla C, Borini A, Fishel S, Griffin D, Campbell A. Plasticity of the human preimplantation embryo: developmental dogmas, variations on themes and self-correction. *Hum Reprod Update* 2021;**27**:848–865.

Coticchio G, Ezoe K, Lagalla C, Shimazaki K, Ohata K, Ninomiya M, Wakabayashi N, Okimura T, Uchiyama K, Kato K *et al.* Perturbations of morphogenesis at the compaction stage affect blastocyst implantation and live birth rates. *Hum Reprod* 2021;**36**: 918–928.

Cruz M, Garrido N, Herrero J, Pérez-Cano I, Muñoz M, Meseguer M. Timing of cell division in human cleavage-stage embryos is linked with blastocyst formation and quality. *Reprod Biomed Online* 2012;**25**:371–381.

Davies S, Christopikou D, Tsorva E, Karagianni T, Mastrominas M, Handyside AH. Altered morphokinetic parameters of embryos identified as aneuploid by single cell array CGH analysis at the 8-cell stage. *Reprod Biomed Online* 2012;**24**:S39–S40.

Davies ST, Christopikou D, Vogiatzi P, Karagianni A, Handyside AH, Mastrominas M. Morphokinetic assessment using time-lapse imaging allows more accurate identification of embryos with multiple rather than single aneuploidies up to the 8-cell stage. Abstract book of the 30th ESHRE Annual Meeting, Munich, Germany, 29 June–2 July 2014. *Hum Reprod* 2014;**29**:i1–i389.

De Cássia Savio Figueira R, Setti AS, De Almeida Ferreira Braga DP, Iaconelli A Jr, Borges E Jr. Blastomere multinucleation: contributing factors and effects on embryo development and clinical outcome. *Hum Fertil (Camb)* 2010;**13**:143–150.

Dal Canto M, Coticchio G, Renzini De Ponti MM, Novara E, Brambillasca PV, Comi F, Fadini R. Cleavage kinetics analysis of human embryos predicts development to blastocyst and implantation. *Reprod Biomed Online* 2012;**25**:474–480.

Del Carmen Nogales M, Bronet F, Basile N, Martínez EM, Liñán A, Rodrigo L, Meseguer M. Type of chromosome abnormality affects embryo morphology dynamics. *Fertil Steril* 2017;**107**:229–235.e2.

Delimitreva SM, Zhivkova RS, Vatev ITS, Toncheva DI. Chromosomal disorders and nuclear and cell destruction in cleaving human embryos. *Int J Dev Biol* 2005;**49**:409–416.

DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–188.

Desai N, Ploskonka S, Goodman LR, Austin C, Goldberg J, Falcone T. Analysis of embryo morphokinetics, multinucleation and cleavage anomalies using continuous time-lapse monitoring in blastocyst transfer cycles. Reprod Biol Endocrinol 2014;**12**:54.

Desai N, Goldberg JM, Austin C, Falcone T. Are cleavage anomalies, multinucleation, or specific cell cycle kinetics observed with time-lapse imaging predictive of embryo developmental capacity or ploidy? *Fertil Steril* 2018;**109**:665–674.

Desai N, Rambhia P. Is there an increase in aneuploidy rate with delayed blastulation, multinucleation or cleavage anomalies? *Fertil Steril* 2016;**106**:e145–e146.

DeUgarte CM, Li M, Surrey M, Danzer H, Hill D, DeCherney AH. Accuracy of FISH analysis in predicting chromosomal status in patients undergoing preimplantation genetic diagnosis. *Fertil Steril* 2008;**90**:1049–1054.

Ebner T, Yaman C, Moser M, Sommergruber M, Pölz W, Tews G. Embryo fragmentation in vitro and its impact on treatment and pregnancy outcome. *Fertil Steril* 2001;**76**:281–285.

Edwards RG, Fishel SB, Cohen J, Fehilly CB, Purdy JM, Slater JM, Steptoe PC, Webster JM. Factors influencing the success of in vitro fertilization for alleviating human infertility. *J In Vitro Fert Embryo Transf* 1984;**1**:3–23.

Esfandiari N, Bunnell ME, Casper RF. Human embryo mosaicism: did we drop the ball on chromosomal testing? *J Assist Reprod Genet* 2016;**33**:1439–1444.

Fan YL, Han SB, Wu LH, Wang YP, Huang GN. Abnormally cleaving embryos are able to produce live births: a time-lapse study. *J Assist Reprod Genet* 2016;**33**:379–385.

Fiorentino F, Bono S, Biricik A, Nuccitelli A, Cotroneo E, Cottone G, Kokocinski F, Michel C-E, Minasi MG, Greco E. Application of next-generation sequencing technology for comprehensive aneuploidy screening of blastocysts in clinical preimplantation genetic screening cycles. *Hum Reprod* 2014;**29**:2802–2813.

Fishel S, Campbell A, Montgomery S, Smith R, Nice L, Duffy S, Jenner L, Berrisford K, Kellam L, Smith R *et al.* Time-lapse imaging algorithms rank human preimplantation embryos according to the probability of live birth. *Reprod Biomed Online* 2018;**37**:304–313.

Fragouli E, Alfarawati S, Spath K, Wells D. Morphological and cytogenetic assessment of cleavage and blastocyst stage embryos. *Mol Hum Reprod* 2014;**20**:117–126.

Fragouli E, Wells D. Aneuploidy in the human blastocyst. *Cytogenet Genome Res* 2011;**133**:149–159.

Franasiak JM, Forman EJ, Hong KH, Werner MD, Upham KM, Treff NR, Scott RT Jr. The nature of aneuploidy with increasing age of the female partner: a review of 15,169 consecutive trophectoderm biopsies evaluated with comprehensive chromosomal screening. *Fertil Steril* 2014;**101**:656–663.e1.

Fréour T, Dessolle L, Lammers J, Lattes S, Barrière P. Comparison of embryo morphokinetics after in vitro fertilization-intracytoplasmic sperm injection in smoking and nonsmoking women. *Fertil Steril* 2013; **99**:1944–1950.

Fujimoto VY, Browne RW, Bloom MS, Sakkas D, Alikani M. Pathogenesis, developmental consequences, and clinical correlations of human embryo fragmentation. *Fertil Steril* 2011;**95**:1197–1204.

Gardner DK, Balaban B. Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': is looking good still important? *Mol Hum Reprod* 2016; **22**:704–718.

Gardner DKS. In vitro culture of human blastocyst. In: Mortimer D (ed). *Towards Reproductive Certainty: Infertility and Genetics Beyond*. Carnforth: Parthenon Press, 1999.

Gazzo E, Pena F, Valdez F, Chung A, Velit M, Ascenzo M, Escudero E. Blastocyst contractions are strongly related with aneuploidy, lower implantation rates, and slow-cleaving embryos: a time lapse study. *JBRA Assist Reprod* 2020;**24**:77–81.

Gleicher N, Orvieto R. Is the hypothesis of preimplantation genetic screening (PGS) still supportable? A review. *J Ovarian Res* 2017;**10**:21.

Goodman LR, Goldberg JM, Falcone T, Desai N. Time-lapse imaging of multinucleated embryos and the association with aneuploidy determined by preimplantation genetic screening. *Fertil Steril* 2016; **106**:e151–e152.

Goossens V, De Rycke M, De Vos A, Staessen C, Michiels A, Verpoest W, Van Steirteghem A, Bertrand C, Liebaers I, Devroey P *et al.* Diagnostic efficiency, embryonic development and clinical

outcome after the biopsy of one or two blastomeres for preimplantation genetic diagnosis. *Hum Reprod* 2008;**23**:481–492.

Grooten WJA, Tseli E, Äng BO, Boersma K, Stålnacke B-M, Gerdle B, Enthoven P. Elaborating on the assessment of the risk of bias in prognostic studies in pain rehabilitation using QUIPS—aspects of interrater agreement. *Diagn Progn Res* 2019;**3**:5.

Hashimoto S, Kato N, Saeki K, Morimoto Y. Selection of high-potential embryos by culture in poly(dimethylsiloxane) microwells and time-lapse imaging. *Fertil Steril* 2012;**97**:332–337.

Hashimoto S, Nakano T, Yamagata K, Inoue M, Morimoto Y, Nakaoka Y. Multinucleation per se is not always sufficient as a marker of abnormality to decide against transferring human embryos. *Fertil Steril* 2016;**106**:133–139.e6.

Hernandez-Nieto C, Lee JA, Slifkin R, Sandler B, Copperman AB, Flisser E. What is the reproductive potential of day 7 euploid embryos? *Hum Reprod* 2019;**34**:1697–1706.

HFEA. Fertility treatment 2019: trends and figures. Accessed at: https://www.hfea.gov.uk/about-us/publications/research-and-data/fertility-treatment-2019-trends-and-figures on May 2022. 2021a.

HFEA. Treatment add-ons with limited evidence. 2021b. Available at: https://www.hfea.gov.uk/treatments/treatment-add-ons/ Accessed on May 2022.

HigginsJPT, ThomasJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane Handbook for Systematic Reviews of Interventions Version 6.2 (updated February 2021). Cochrane. 2021.

Ho JR, Arrach N, Rhodes-Long K, Salem W, McGinnis LK, Chung K, Bendikson KA, Paulson RJ, Ahmady A. Blastulation timing is associated with differential mitochondrial content in euploid embryos. *J Assist Reprod Genet* 2018;**35**:711–720.

Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can predict embryo ploidy status based on time-lapse data. *Reprod Biol Endocrinol* 2021;**19**:185.

Kaing A, Kroener LL, Tassin R, Li M, Liu L, Buyalos R, Hubert G, Shamonki M. Earlier day of blastocyst development is predictive of embryonic euploidy across all ages: essential data for physician decision-making and counseling patients. *J Assist Reprod Genet* 2018;**35**:119–125.

Kalatova B, Jesenska R, Hlinka D, Dudas M. Tripolar mitosis in human cells and embryos: occurrence, pathophysiology and medical implications. *Acta Histochem* 2015;**117**:111–125.

Kemper JM, Liu Y, Afnan M, Hammond ER, Morbeck DE, Mol BWJ. Should we look for a low-grade threshold for blastocyst transfer? A scoping review. *Reprod Biomed Online* 2021;**42**:709–716.

Kim S-G, Kim Y-Y, Park J-Y, Kwak S-J, Yoo C-S, Park I-H, Sun H-G, Kim J-W, Lee K-H, Park H-D *et al*. Early fragment removal on in vitro fertilization day 2 significantly improves the subsequent development and clinical outcomes of fragmented human embryos. *Clin Exp Reprod Med* 2018;**45**:122–128.

Kimelman D, Confino R, Okeigwe I, Lambe-Steinmiller J, Confino E, Shulman LP, Zhang JX, Pavone ME. Assessing the impact of delayed blastulation using time lapse morphokinetics and preimplantation genetic testing in an IVF patient population. *J Assist Reprod Genet* 2019;**36**:1561–1569.

Kirkegaard K, Kesmodel US, Hindkjær JJ, Ingerslev HJ. Time-lapse parameters as predictors of blastocyst development and pregnancy outcome in embryos from good prognosis patients: a prospective cohort study. *Hum Reprod* 2013a;**28**:2643–2651.

Kirkegaard K, Hindkjaer JJ, Ingerslev HJ. Effect of oxygen concentration on human embryo development evaluated by time-lapse monitoring. *Fertil Steril* 2013b;**99**:738–744.e4.

Kligman I, Benadiva C, Alikani M, Munne S. The presence of multinucleated blastomeres in human embryos is correlated with chromosomal abnormalities. *Hum Reprod* 1996;**11**:1492–1498.

Kramer YG, Kofinas JD, Melzer K, Noyes N, McCaffrey C, Buldo-Licciardi J, McCulloh DH, Grifo JA. Assessing morphokinetic parameters via time lapse microscopy (TLM) to predict euploidy: are aneuploidy risk classification models universal? *J Assist Reprod Genet* 2014;**31**:1231–1242.

Lagalla C, Tarozzi N, Sciajno R, Nadalini M, Di Santo A, Distratis V, Wells D, Borini A. Embryos with cell division aberrations monitored by time-lapse imaging in a PGS program: are they able to develop into euploid blastocysts? *Hum Reprod* 2015;**30**.

Lagalla C, Tarozzi N, Sciajno R, Wells D, Di Santo M, Nadalini M, Distratis V, Borini A. Embryos with morphokinetic abnormalities may develop into euploid blastocysts. *Reprod Biomed Online* 2017;**34**:137–146.

Leary C, Leese HJ, Sturmey RG. Human embryos from overweight and obese women display phenotypic and metabolic abnormalities. *Hum Reprod* 2015;**30**:122–132.

Lee CI, Chen CH, Huang CC, Cheng EH, Chen HH, Ho ST, Lin PY, Lee MS, Lee TH. Embryo morphokinetics is potentially associated with clinical outcomes of single-embryo transfers in preimplantation genetic testing for aneuploidy cycles. *Reprod Biomed Online* 2019;**39**:569–579.

Lemmen JG, Agerholm I, Ziebe S. Kinetic markers of human embryo quality using time-lapse recordings of IVF/ICSI-fertilized oocytes. *Reprod Biomed Online* 2008;**17**:385–391.

Li F, Urich M, Ayers JW, Abuzeid MI, Khan I. Euploidy is not predictable by time lapse imaging system with either blastomere or trophectoderm biopsies. *Fertil Steril* 2015;**104**:e314–e315.

Liu Y, Chapple V, Roberts P, Matson P. Prevalence, consequence, and significance of reverse cleavage by human embryos viewed with the use of the embryoscope time-lapse video system. *Fertil Steril* 2014;**102**:1295–1300.e2.

Los FJ, Van Opstal D, van den Berg C. The development of cytogenetically normal, abnormal and mosaic embryos: a theoretical model. *Hum Reprod Update* 2004;**10**:79–94.

Magata F, Ideta A, Okubo H, Matsuda F, Urakawa M, Oono Y. Growth potential of bovine embryos presenting abnormal cleavage observed through time lapse cinematography. *Theriogenology* 2019;**133**:119–124.

Magli MC, Gianaroli L, Ferraretti AP. Chromosomal abnormalities in embryos. *Mol Cell Endocrinol* 2001;**183**:S29–S34.

Magli MC, Gianaroli L, Ferraretti AP, Lappi M, Ruberti A, Farfalli V. Embryo morphology and development are dependent on the chromosomal complement. *Fertil Steril* 2007;**87**:534–541.

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;**22**:719–748.

Marcos J, Pérez-Albalá S, Mifsud A, Molla M, Landeras J, Meseguer M. Collapse of blastocysts is strongly related to lower implantation success: a time-lapse study. *Hum Reprod* 2015;**30**:2501–2508.

Martin A, Rodrigo L, Beltran D, Meseguer M, Rubio C, Mercader A, de Los Santos MJ. The morphokinetic signature of mosaic

embryos: evidence in support of their own genetic identity. *Fertil Steril* 2021;**116**:165–173.

Mazur P, Nagornyy V, Mykytenko D, Semeniuk L, Zukin V. Multinucleation and aneuploidy, time-lapse imaging and aCGH may shed the light on causation. *Fertil Steril* 2013;**100**:S68.

McCoy RC. Mosaicism in preimplantation human embryos: when chromosomal abnormalities are the norm. *Trends Genet* 2017;**33**: 448–463.

McCoy RC, Newnham LJ, Ottolini CS, Hoffmann ER, Chatzimeletiou K, Cornejo OE, Zhan Q, Zaninovic N, Rosenwaks Z, Petrov DA *et al.* Tripolar chromosome segregation drives the association between maternal genotype at variants spanning PLK4 and aneuploidy in human preimplantation embryos. *Hum Mol Genet* 2018; **27**:2573–2585.

McDaniel KE, Awadalla MS, McGinnis LK, Ahmady A. Transfer the best and biopsy the rest? Blastocyst euploidy rates differ by morphology and day of biopsy. *Arch Gynecol Obstet* 2021;**303**: 249–258.

Melzer KE, Noyes N, Hodes-Wertz B, McCulloh D, Munne S, Grifo JA. How well do morphokinetic (MK) parameters and time-lapse microscopy (TLM) predict euploidy? A pilot study of TLM with trophectoderm (TE) biopsy with array comparative genomic hybridization (aCGH). *Fertil Steril* 2013;**100**:S209.

Meriano J, Clark C, Cadesky K, Laskin CA. Binucleated and micronucleated blastomeres in embryos derived from human assisted reproduction cycles. *Reprod Biomed Online* 2004;**9**:511–520.

Milewski R, Milewska AJ, Kuczyńska A, Stankiewicz B, Kuczyński W. Do morphokinetic data sets inform pregnancy potential? *J Assist Reprod Genet* 2016;**33**:357–365.

Minasi MG, Colasante A, Riccio T, Ruberti A, Casciani V, Scarselli F, Spinella F, Fiorentino F, Varricchio MT, Greco E. Correlation between aneuploidy, standard morphology evaluation and morphokinetic development in 1730 biopsied blastocysts: a consecutive case series study. *Hum Reprod* 2016;**31**:2245–2254.

Moayeri SE, Allen RB, Brewster WR, Kim MH, Porto M, Werlin LB. Day-3 embryo morphology predicts euploidy among older subjects. *Fertil Steril* 2008;**89**:118–123.

Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 2009;**62**:1006–1012.

Mumusoglu S, Yarali I, Bozdag G, Ozdemir P, Polat M, Sokmensuer LK, Yarali H. Time-lapse morphokinetic assessment has low to moderate ability to predict euploidy when patient- and ovarian stimulation-related factors are taken into account with the use of clustered data analysis. Fertil Steril 2017;**107**:413–421.e4.

Munné S. Chromosome abnormalities and their relationship to morphology and development of human embryos. *Reprod Biomed Online* 2006;**12**:234–253.

Munné S, Blazek J, Large M, Martinez-Ortiz PA, Nisson H, Liu E, Tarozzi N, Borini A, Becker A, Zhang J *et al.* Detailed investigation into the cytogenetic constitution and pregnancy outcome of replacing mosaic blastocysts detected with the use of high-resolution next-generation sequencing. *Fertil Steril* 2017;**108**:62–71.e8.

Munne S, Cohen J. Chromosome abnormalities in human embryos. *Hum Reprod Update* 1998;**4**:842–855.

Munné S, Kaplan B, Frattarelli JL, Child T, Nakhuda G, Shamma FN, Silverberg K, Kalista T, Handyside AH, Katz-Jaffe M *et al.*; STAR Study Group. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. *Fertil Steril* 2019;**112**:1071–1079.e7.

Munoz EO, Perez M, Taboas E, Meseguer M, Aguilar J. Morphology dynamics of multinucleation affects embryo chromosome content. Abstract book of the 30th ESHRE Annual Meeting, Munich, Germany, 29 June–2 July 2014. *Hum Reprod* 2014;**29**:i1–i389.

Muñoz M, Cruz M, Humaidan P, Garrido N, Pérez-Cano I, Meseguer M. The type of GnRH analogue used during controlled ovarian stimulation influences early embryo developmental kinetics: a time-lapse study. *Eur J Obstet Gynecol Reprod Biol* 2013;**168**:167–172.

Neal SA, Sun L, Jalas C, Morin SJ, Molinaro TA, Scott RT. When next-generation sequencing-based preimplantation genetic testing for aneuploidy (PGT-A) yields an inconclusive report: diagnostic results and clinical outcomes after re biopsy. *J Assist Reprod Genet* 2019;**36**:2103–2109.

Niimura S. Time-lapse videomicrographic analyses of contractions in mouse blastocysts. *J Reprod Dev* 2003;**49**:413–423.

Ozbek IY, Mumusoglu S, Polat M, Bozdag G, Sokmensuer LK, Yarali H. Comparison of single euploid blastocyst transfer cycle outcome derived from embryos with normal or abnormal cleavage patterns. *Reprod Biomed Online* 2021;**42**:892–900.

Patel DV, Shah PB, Kotdawala AP, Herrero J, Rubio I, Banker MR. Morphokinetic behavior of euploid and aneuploid embryos analyzed by time-lapse in embryoscope. *J Hum Reprod Sci* 2016;**9**: 112–118.

Petersen BM, Boel M, Montag M, Gardner DK. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3. *Hum Reprod* 2016;**31**:2231–2244.

Pickering SJ, Taylor A, Johnson MH, Braude PR. An analysis of multinucleated blastomere formation in human embryos. *Hum Reprod* 1995;**10**:1912–1922.

Pribenszky C, Nilselid AM, Montag M. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. *Reprod Biomed Online* 2017;**35**:511–520.

Puissant F, Van Rysselberge M, Barlow P, Deweze J, Leroy F. Embryo scoring as a prognostic tool in IVF treatment. *Hum Reprod* 1987;**2**: 705–708.

Rienzi L, Capalbo A, Stoppa M, Romano S, Maggiulli R, Albricci L, Scarica C, Farcomeni A, Vajta G, Ubaldi FM. No evidence of association between blastocyst aneuploidy and morphokinetic assessment in a selected population of poor-prognosis patients: a longitudinal cohort study. *Reprod Biomed Online* 2015;**30**:57–66.

Rienzi L, Bono S, Capalbo L, Spizzichino C, Rubio FM, Fiorentino Ubaldi F. Human embryo aneuploidues cannot be predicted by morphokinetic assessement during the first phases of in vitro culture. Abstract Book. *Hum Reprod.* 2013;**28**:i149–i206.

Riley RD, van der Windt D, Croft P, Moons KG. *Prognosis Research in Healthcare: Concepts, Methods, and Impact.* Oxford: Oxford University Press, 2019.

Royen EV, Mangelschots K, Vercruyssen M, Neubourg DD, Valkenburg M, Ryckaert G, Gerris J. Multinucleation in cleavage stage embryos. *Hum Reprod* 2003;**18**:1062–1069.

Ruangvutilert P, Delhanty JD, Rodeck CH, Harper JC. Relative efficiency of FISH on metaphase and interphase nuclei from non-mosaic trisomic or triploid fibroblast cultures. *Prenat Diagn* 2000;**20**:159–162.

Rubio I, Kuhlmann R, Agerholm I, Kirk J, Herrero J, Escribá M-J, Bellver J, Meseguer M. Limited implantation success of direct-cleaved human zygotes: a time-lapse study. *Fertil Steril* 2012;**98**:1458–1463.

Schünemann H, Brożek J, Guyatt G, Oxman A (eds). *GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations*. The GRADE Working Group, Cochrane Collaboration, 2013.

Scott LF, Kloos B, Davies D. 2010. The origin and consequences of Day-2 multinucleation of human embryos. Conference: 26th Annual Meeting of the European Society of Human Reproduction and Embryology, ESHRE. Rome Italy. Conference Publication: (var.pagings). 25 (SUPPL. 1) (pp i195-i196), 2010. Date of Publication: June 2010. *Hum Reprod* 2010:**25**:i170–i210.

Scriven PN, Bossuyt PM. Diagnostic accuracy: theoretical models for preimplantation genetic testing of a single nucleus using the fluorescence in situ hybridization technique. *Hum Reprod* 2010;**25**: 2622–2628.

Staessen C, Platteau P, Van Assche E, Michiels A, Tournaye H, Camus M, Devroey P, Liebaers I, Van Steirteghem A. Comparison of blastocyst transfer with or without preimplantation genetic diagnosis for aneuploidy screening in couples with advanced maternal age: a prospective randomized controlled trial. *Hum Reprod* 2004; **19**:2849–2858.

Stroup DF. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;**283**:2008–2012.

Su Y, Li J-J, Wang C, Haddad G, Wang W-H. Aneuploidy analysis in day 7 human blastocysts produced by in vitro fertilization. *Reprod Biol Endocrinol* 2016;**14**:20–27.

Tarín JJ, Conaghan J, Winston RM, Handyside AH. Human embryo biopsy on the 2nd day after insemination for preimplantation diagnosis: removal of a quarter of embryo retards cleavage. *Fertil Steril* 1992;**58**:970–976.

Taylor-Rowan M, Edwards S, Noel-Storr AH, McCleery J, Myint PK, Soiza R, Stewart C, Loke YK, Quinn TJ. Anticholinergic burden (prognostic factor) for prediction of dementia or cognitive decline in older adults with no known cognitive syndrome. *Cochrane Database Syst Rev* 2021;**5**:CD013540.

Theobald R, SenGupta S, Harper J. The status of preimplantation genetic testing in the UK and USA. *Hum Reprod* 2020;**35**:986–998.

Tiegs AW, Sun L, Patounakis G, Scott RT. Worth the wait? Day 7 blastocysts have lower euploidy rates but similar sustained implantation rates as Day 5 and Day 6 blastocysts. Human Reproduction 2019;**34**:1632–1639.

Tiegs AW, Tao X, Zhan Y, Whitehead C, Kim J, Hanson B, Osman E, Kim TJ, Patounakis G, Gutmann J *et al*. A multicenter, prospective, blinded, nonselection study evaluating the predictive value of an aneuploid diagnosis using a targeted next-generation sequencing-based preimplantation genetic testing for aneuploidy assay and impact of biopsy. *Fertil Steril* 2021;**115**:627–637.

Vera-Rodriguez M, Chavez SL, Rubio C, Reijo Pera RA, Simon C. Prediction model for aneuploidy in early human embryo development revealed by single-cell analysis. *Nat Commun* 2015;**6**:7601.

Victor AR, Griffin DK, Brake AJ, Tyndall JC, Murphy AE, Lepkowsky LT, Lal A, Zouves CG, Barnes FL, McCoy RC *et al*. Assessment of aneuploidy concordance between clinical trophectoderm biopsy and blastocyst. *Hum Reprod* 2019;**34**:181–192.

Vinals Gonzalez X, Odia R, Cawood S, Gaunt M, Saab W, Seshadri S, Serhal P. Contraction behaviour reduces embryo competence in high-quality euploid blastocysts. *J Assist Reprod Genet* 2018;**35**: 1509–1517.

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007;**370**:1453–1457.

Watson AJ, Natale DR, Barcroft LC. Molecular regulation of blastocyst formation. *Anim Reprod Sci* 2004;**82–83**:583–592.

Werland H, VerMilyea M, Silverberg K. Culturing embryos to day 7. A viable option for IVF patients? *Fertil Steril* 2017;**108**:e161–e162.

Whitney J, Anderson R, Schiewe M. Day 7 blastocyst euploidy and implantation rates warrant implentation for all programs using preimplantation genetic screening (PGS). *Fertil Steril* 2016;**106**:e146.

Wong CC, Loewke KE, Bossert NL, Behr B, De Jonge CJ, Baer TM, Pera RAR. Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat Biotechnol* 2010;**28**:1115–1121.

Yang Z, Zhang J, Salem SA, Liu X, Kuang Y, Salem RD, Liu J. Selection of competent blastocysts for transfer by combining time-lapse monitoring and array CGH testing for patients undergoing preimplantation genetic screening: a prospective study with sibling oocytes. *BMC Med Genomics* 2014;**7**:38.

Yilmaz A, Zhang L, Zhang XY, Son WY, Holzer H, Ao A. Chromosomal complement and clinical relevance of multinucleated embryos in PGD and PGS cycles. *Reprod Biomed Online* 2014;**28**: 380–387.

Zaninovic N, Irani M, Meseguer M. Assessment of embryo morphology and developmental dynamics by time-lapse microscopy: is there a relation to implantation and ploidy? *Fertil Steril* 2017;**108**: 722–729.

Zhan Q, Ye Z, Clarke R, Rosenwaks Z, Zaninovic N. Direct unequal cleavages: embryo developmental competence, genetic constitution and clinical outcome. *PLoS One* 2016;**11**:e0166398.

Zhang J, Tao W, Liu H, Yu G, Li M, Ma S, Wu K. Morphokinetic parameters from a time-lapse monitoring system cannot accurately predict the ploidy of embryos. *J Assist Reprod Genet* 2017;**34**: 1173–1178.

Ziebe S, Lundin K, Loft A, Bergh C, Andersen AN, Selleskog U, Nielsen D, Grondahl C, Kim H, Arce JC *et al*. FISH analysis for chromosomes 13, 16, 18, 21, 22, X and Y in all blastomeres of IVF pre-embryos from 144 randomly selected donated human oocytes and impact on pre-embryo morphology. *Hum Reprod* 2003;**18**:2575–2581.

Ziebe S, Petersen K, Lindenberg S, Andersen AG, Gabrielsen A, Andersen AN. Embryo morphology or cleavage stage: how to select the best embryos for transfer after in-vitro fertilization. *Hum Reprod* 1997;**12**:1545–1549.

CHAPTER 3: A COMPARISON OF 12 MACHINE LEARNING MODELS DEVELOPED TO PREDICT PLOIDY, USING A MORPHOKINETIC META-DATASET OF 8147 EMBRYOS

Thomas Bamford, Christina Easter, Sue Montgomery, Rachel Smith, Rima K. Dhillon-Smith, Amy Barrie, Alison Campbell, Arri Coomarasamy

Manuscript published in Human Reproduction, Vol. 38, No. 4, pp. 569-581, 24th February 2023

https://doi.org/10.1093/humrep/dead034

My role in this publication is as follows: I collected and cleaned all the data, built and created each of the 12 machine learning models. I learnt coding and the use of the University of Birmingham supercomputer, BlueBEAR. Following this I then conducted all analyses as part of the validation study and wrote the entirety of the manuscript.

Amendments: where this article mentioned ‘traditional statistics’ it is referring to logistic regression. Short protocol is synonymous with short antagonist protocol.

Human Reproduction, Vol.38, No.4, pp. 569–581, 2023
Advance Access Publication on February 24, 2023 https://doi.org/10.1093/humrep/dead034

human reproduction

ORIGINAL ARTICLE *Embryology*

A comparison of 12 machine learning models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos

Thomas Bamford [1,*], Christina Easter [1], Sue Montgomery [2], Rachel Smith [2], Rima K. Dhillon-Smith [1], Amy Barrie [2], Alison Campbell [2], and Arri Coomarasamy [1]

[1]Tommy's National Centre for Miscarriage Research, Institute of Metabolism and Systems Research, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, UK [2]Care Fertility Headquarters, Nottingham, UK

*Correspondence address. CARE Fertility Manchester, 108-112 Daisy Bank Rd, Manchester M14 5QH, UK. Tel: +44-01612493040; E-mail: tom.bamford@carefertility.com https://orcid.org/0000-0001-9048-3912

Submitted on November 10, 2022; resubmitted on January 31, 2023; editorial decision on February 8, 2023

STUDY QUESTION: Are machine learning methods superior to traditional statistics in predicting blastocyst ploidy status using morphokinetic and clinical biodata?

SUMMARY ANSWER: Mixed effects logistic regression performed better than all machine learning methods for ploidy prediction using our dataset of 8147 embryos.

WHAT IS KNOWN ALREADY: Morphokinetic timings have been demonstrated to be delayed in aneuploid embryos. Machine learning and statistical models are increasingly being built, however, until now they have been limited by data insufficiency.

STUDY DESIGN, SIZE, DURATION: This is a multicentre cohort study. Data were obtained from 8147 biopsied blastocysts from 1725 patients, treated from 2012 to 2020.

PARTICIPANTS/MATERIALS, SETTING, METHODS: All embryos were cultured in a time-lapse system at nine IVF clinics in the UK. A total of 3004 euploid embryos and 5023 aneuploid embryos were included in the final verified dataset. We developed a total of 12 models using four different approaches: mixed effects multivariable logistic regression, random forest classifiers, extreme gradient boosting, and deep learning. For each of the four algorithms, two models were created, the first consisting of 22 covariates using 8027 embryos (Dataset 1) and the second, a dataset of 2373 embryos and 26 covariates (Dataset 2). Four final models were created by switching the target outcome from euploid to aneuploid for each algorithm (Dataset 1). Models were validated using internal–external cross-validation and external validation.

MAIN RESULTS AND THE ROLE OF CHANCE: All morphokinetic variables were significantly delayed in aneuploid embryos. The likelihood of euploidy was significantly increased the more expanded the blastocyst ($P < 0.001$) and the better the trophectoderm grade ($P < 0.01$). Univariable analysis showed no association with ploidy status for morula or cleavage stage fragmentation, morula grade, fertilization method, sperm concentration, or progressive motility. Male age did not correlate with the percentage of euploid embryos when stratified for female age. Multinucleation at the two-cell or four-cell stage was not associated with ploidy status. The best-performing model was logistic regression built using the larger dataset with 22 predictors (F1 score 0.59 for predicting euploidy; F1 score 0.77 for predicting aneuploidy; AUC 0.71; 95% CI 0.67–0.73). The best-performing models using the algorithms from random forest, extreme gradient boosting, and deep learning achieved an AUC of 0.68, 0.63, and 0.63, respectively. When using only morphokinetic predictors the AUC was 0.61 for predicting ploidy status, whereas a model incorporating only embryo grading was unable to discriminate aneuploid embryos (AUC = 0.52). The ploidy prediction model's performance improved with increasing age of the egg provider.

LIMITATIONS, REASONS FOR CAUTION: The models have not been validated in a prospective study design or yet been used to determine whether they improve clinical outcomes

WIDER IMPLICATIONS OF THE FINDINGS: This model may aid decision-making, particularly where pre-implantation genetic testing for aneuploidy is not permitted or for prioritizing embryos for biopsy.

STUDY FUNDING/COMPETING INTEREST(S): No specific funding was sought for this study; university funds supported the first author. A.Ca. is a minor shareholder of participating centres.

TRIAL REGISTRATION NUMBER: N/A.

Key words: morphokinetcs / ploidy / model / artificial intelligence / machine learning / euploid / aneuploid

Introduction

A time-lapse incubation system (TLS) allows the annotation of an embryo's development milestones, termed morphokinetics, such as the time to formation of a blastocyst. Embryo development can be compared to outcome variables, such as ploidy status or live birth. Embryos are selected for transfer that display specific development patterns using statistical models. Notably, a Cochrane review concluded that there was insufficient good quality evidence of differences in live birth between TLS and conventional morphological assessment (Armstrong *et al.*, 2019). In contrast, potential benefits have been described elsewhere, such as a reduced miscarriage rate; therefore, despite controversies a TLS has become commonplace in many IVF laboratories worldwide (Pribenszky *et al.*, 2017).

It has been hypothesized that delayed cleavage is a consequence of chromosomal abnormalities (Campbell *et al.*, 2013a,b). Therefore, studies started to investigate whether these kinetic markers could be harnessed as a non-invasive method of embryo ploidy assessment. This has been evidenced by a recent meta-analysis of 58 studies, reporting 10 morphokinetic variables to be significantly delayed in aneuploid embryos (Bamford *et al.*, 2022).

This study focuses on ploidy status over live birth for several reasons. Firstly, the 'ground truth' data can be misleading in live birth data. For instance, if an embryo was transferred in an inadequately prepared endometrium the model learns that this embryo is nonviable owing to it resulting in a negative pregnancy test. Secondly, pre-implantation genetic testing for aneuploidy (PGT-A) is increasingly being offered to patients worldwide, for example, the USA perform 27% PGT-A cycles compared to <2% in the UK (Theobald *et al.*, 2020). This is despite randomized controlled trials demonstrating no improvement in live birth rates, although possibly a reduced miscarriage rate and shortened time to pregnancy in older patients (Neal *et al.*, 2018; Munné *et al.*, 2019; Cornelisse *et al.*, 2020). While the pregnancy rate per cycle may not increase, the clinical pregnancy rate for a biopsied euploid embryo is significantly higher than that of an untested embryo. This has been demonstrated by Sanders *et al.*, (2021), who analysed data from the Human Fertilisation and Embryology Authority (HFEA) for 2016–2018. This incorporated 2464 PGT-A cycles, and all age groups showed a higher live birth rate when PGT-A was used; this was significantly higher (5- to 10-fold) for those >37 years of age. For instance, the live birth rate per euploid embryo transfer for 38–39-year-olds was 42% versus 17% for an untested embryo. It has also been demonstrated that the cumulative live birth rate after three euploid embryo transfers reaches 93% (Pirtea *et al.*, 2021). Furthermore, non-selection studies have frequently found that the live birth rate from aneuploid embryos is 0%, making selecting out these embryos seem advantageous (Tiegs *et al.*, 2020). The use of PGT-A is undoubtedly increasing worldwide, therefore, regardless of its effectiveness, finding non-invasive alternatives remains an important venture. This is particularly important for countries where embryo biopsy remains prohibited and for patients who may have ethical or religious objections to the procedure. Finally, PGT-A is expensive, therefore creating an alternative, less invasive method may make this technology more accessible to patients.

Several studies have now used morphokinetic models to predict ploidy status with varied success (Campbell *et al.*, 2013a; Chawla *et al.*, 2015; Uyar *et al.*, 2015; Del Carmen Nogales *et al.*, 2017; Mumusoglu *et al.*, 2017; Desai *et al.*, 2018; Huang *et al.*, 2021). A significant limitation of this published work is the limited sample size for the training and testing of the models and heterogeneity in the validation procedures. That said, we now have increasing access to 'big data', particularly with the increased use of artificial intelligence (AI) in reproductive medicine. With this in mind, it is therefore justified to explore this area further with a large genetic dataset. This study aims to compare model performance metrics for different AI algorithms built to predict ploidy status using morphokinetics, morphological, and clinical data. We will determine whether newer, more complex algorithms (without image analysis) offer improved results to traditional statistics. This is important for clinics without access to expensive blastocyst image analysis software and allows optimization of the use of time lapse for embryo selection.

Materials and methods

Data for this cohort study were obtained from the treatment of 1725 patients from 2012 to 2020 at 9 IVF clinics in the UK, using 8147 biopsied blastocysts. The participating centres are part of a private fertility group, each conforming to the same laboratory practices. There were no major changes in laboratory practices during the study period apart from a change in the culture medium used. Detailed comparisons were undertaken to ensure there was no difference in the median timing for morphokinetic variables between old and new media. Key performance indicators were at least maintained during this study period and existing morphokinetic embryo selection models remained equally effective. This study did not require ethical approval as it utilized anonymized, numerical data acquired during routine, validated and HFEA-licenced practices. The data analyses were approved by the participating clinics research and innovation board, and it was decided that consent was not necessary for participation in this retrospective data analysis as patient care and fate of the embryos was unaffected. The study is reported in accordance with Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

(TRIPOD) guidelines (Supplementary Table SI) (Collins *et al.*, 2015, 2021). Figure 1 provides an overview of the study methodology.

Participants

The patients selected for PGT-A were a heterogenous group of patients who were recommended PGT-A largely for advanced maternal age, recurrent implantation failure (>2 failed embryo transfers), recurrent miscarriage (>2 spontaneous miscarriages), or to shorten the time to pregnancy. All PGT-A patients were included and mosaic embryos were excluded from the initial modelling. Patients underwent pituitary suppression using either the long GnRH agonist or short antagonist protocol. Ovarian stimulation dose was determined by the patient's anti-Müllerian hormone, antral follicle count, BMI, and previous response. Transvaginal oocyte retrieval was performed under sedation 36 h after the trigger injection (hCG or agonist) when three follicles reached 17 mm or more.

Laboratory procedures and trophectoderm biopsy

Aspirated oocytes were inseminated by ICSI or IVF. All embryos were cultured in an EmbryoScope TLS (Vitrolife; Frölunda, Sweden) at 37°C, 6–6.5% CO_2, 5% O_2, and 88.5–89% N_2. The wells of the Embryoscope slide (Vitrolife; Frölunda, Sweden) were filled with culture medium, Global total LP (Cooper Surgical; Trumbull, CT, USA) covered with LifeGuard mineral oil (Cooper Surgical; Trumbull, CT, USA). Every 10 min microscopic images were acquired of the embryos through seven multifocal planes.

All blastocysts were assessed for suitability for biopsy on Day 5 or 6. Laser pulses were used to facilitate trophectoderm herniation, 5–10 cells were then aspirated as described in a previous publication (Campbell *et al.*, 2013a). Biopsied embryos were vitrified while awaiting results.

Outcome

Euploidy was considered the primary target outcome. The genetic platform used for determining the outcome was array comparative genomic hybridization (aCGH) in 367 (5%) and next-generation sequencing in 7660 (95%) blastocysts. PGT-A samples with <20% aneuploidy were classed as euploid, 20–80% mosaic, and >80% aneuploid. The impact of switching the target variable to aneuploidy was explored. There was no blinding of the outcome necessary.

Predictors

A systematic review and meta-analysis were undertaken to determine the variables with signal for model development (Bamford *et al.*, 2022). Morphokinetic parameters were manually annotated by an embryologist. All annotators were trained in house and followed a strict annotation procedure, supported by regular monitoring and quality assurance procedures validated elsewhere (Barrie, 2021). Annotations included: time (from insemination) to second polar body extrusion (tPB2), pro-nuclear fade (tPNf), time to n cells (tn), time to start of compaction (tSC), time to morula (tM), time to start of blastulation (tSB), time to full blastocyst (tB), time to expanded blastocyst (tEB), and time to hatching blastocyst (tHB). Annotation definitions are described elsewhere (Campbell *et al.*, 2013a). Morphokinetics were expressed as absolute times rather than interim times.

The presence of multinucleation was assessed at the two or four-cell stage. Fragmentation was categorized using the following thresholds, 0–10%, 10–20%, 20–50%, and >50%. This was assessed at the cleavage stage on Day 3 (68 h). The morula was graded as 'M1' if fully compacted and 'M2' where partial exclusion of cells was observed. Blastocysts were classified according to the Gardner criteria on Day 5, considering the degree of expansion, trophectoderm and inner cell mass grade as individual predictors (Gardner, 1999). Sperm concentration and progressive motility were recorded as per World Health Organization guidelines (WHO, 2021). All other clinical and embryological predictors were recorded as part of routine workflows in each clinic. The assessment of predictors was performed whilst blinding the PGT-A results, information about the other predictors could not be blinded.

Preparing data and handling missing values

tEB and tHB were excluded from all analyses as embryos may have been biopsied prior to reaching this development stage and, similarly, not all embryos will display nine cells, therefore t9 was excluded. An attempt was made to fill missing values using manual data entry. Multiple imputation of missing values up to a proportion of 20% has been validated, beyond this threshold it is less certain whether this method continues to reliably impute (Musil *et al.*, 2002). In the final datasets, no variables required imputation as all those included had <5% missing; they were therefore considered to be missing at random (Peng *et al.*, 2006; Tabachnick *et al.*, 2007; Jakobsen *et al.*, 2017). In the AI algorithms, these values were imputed using the median value for that variable (Jadhav *et al.*, 2019). The data were transformed using normalization. Sensitivity analyses determined whether imputation using k-nearest neighbours rather than median, or whether the use of standardization rather than normalization, improved model performance (Triguero *et al.*, 2019).

Routine recording of certain variables in the databases increased over time, therefore we aimed to create two more complete datasets. Dataset 1 (2012–2020) included all morphokinetic variables and all other predictors were where <20% were missing. This study sought to construct a second, more complete, dataset using embryos from 2017 onwards.

Further analyses

In addition to the primary analyses, we investigated whether the presence of multinucleation at the two-cell or four-cell stage was a predictor of ploidy status. The impact of male ageing on blastocyst ploidy status when stratified for female age was investigated. A *post hoc* analysis was conducted to determine whether mosaic embryos could be identified from euploid embryos. Additionally, a morphological model was created, using only the components of the Gardner classification, to determine whether morphokinetics offers anything over and above embryo grading. Finally, we also explored whether adjusting the morphokinetic timings to tPNf = 0 or using the interval times as model features would improve the model.

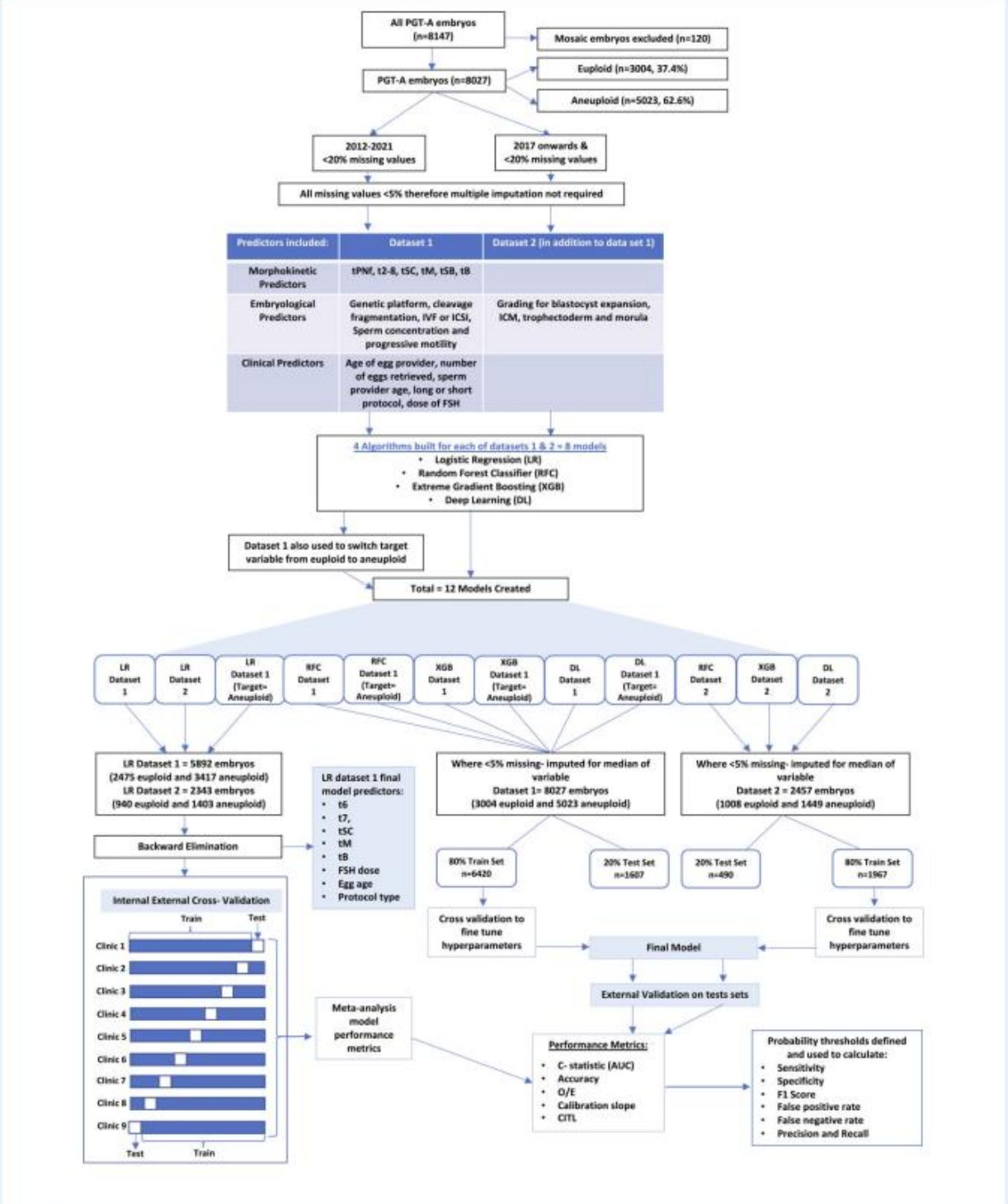


Figure 1. Overview of models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos. PGT-A: pre-implantation genetic testing for aneuploidy; tPNf: time to pronuclear fade; tn: time to n cells; tSC: time to start of compaction; tM: time to formation of Morula; tSB: time to start of blastulation; tB: time to formation of blastocyst; O/E: Observed to Expected ratio; CITL: calibration in the large.

Sample size calculations

Based on 26 potential predictive variables being entered into the model and a prevalence of 40% for euploidy, a minimum sample size of 1672 with 669 events was required (Riley *et al.*, 2019). This calculation was based on a shrinkage factor of 1 and an R^2 value of 0.15 (Riley *et al.*, 2019).

Training and validation of machine learning models

Logistic regression

Logistic regression (LR) is the traditional statistical approach to modelling to which we will be comparing the AI methods. To account for repeated IVF cycles within participants and the clustering of embryos per patient, this study used mixed effects LR using the patient identifier as the random intercept, allowing the within-cluster variability to be accounted for. Predictor distributions were analysed and tested for collinearity. Variables were selected for inclusion in the model using backward elimination with a *P*-value of <0.157, a methodology validated elsewhere (Sauerbrei, 1999; Chowdhury and Turin, 2020). No attempt was made to fit a new model using only features with statistical significance as this can lead to overfitting and bias.

AI models

Three additional models were built using the algorithms: random forest classifier (RFC), extreme gradient boosting (XGBoost), and deep learning (DL). An outline of the terminology can be found in Supplementary Table SII. Each of these models was trained using 80% of the data (n = 6420 in Dataset 1; n = 1967 in Dataset 2), and stratification was used to ensure an equal proportion of euploid embryos in each of the test and training sets. Categorical variables were converted using one hot encoding (Potdar *et al.*, 2017). All hyperparameters available in each algorithm's library were tuned for using k-fold cross-validation. Feature selection and an ensemble method using principal component analysis was attempted for each of the AI models; this weakened the performance and therefore all variables were included in all AI modelling. RFC is a classification algorithm made up of many decision trees using a bootstrapped dataset, constructed using substantiated methods (Breiman, 2001; Cutler *et al.*, 2012; Ayyadevara, 2018). The forest chooses the classification with the majority of the 'votes' given by the individual trees. XGBoost has similarities to RFC, however, the trees are added sequentially and fit to correct the prediction errors made by prior models (Putatunda and Rama, 2018). Finally, this study tested a DL algorithm, known as an artificial neural network, using a multi-layer perceptron model using the sequential technique. This comprises node layers, containing an input layer, one or more hidden layers (where complex non-linear functions are applied with different weights), and an output layer.

Model validation and performance metrics

The LR model was validated using internal–external cross-validation (IECV). This iteratively selected n-1 clinics from the total nine clinics, and the prognostic model was developed within this subset of clinics, leaving the remaining clinic for external validation (Steyerberg and Harrell, 2016). The key performance statistics were meta-analysed using random effects across nine clinics forming the external datasets.

The AI models were optimized and adapted through each fold of cross-validation using 80% of the datasets; because we used cross-validation to tune hyperparameters, we conducted external validation on the remaining 20% of the data. Performance metrics included: AUC and accuracy. Using a defined threshold of predicted probability, we tested the ability of each model to correctly classify embryos. From this, precision, false positive rate, false negative rate, sensitivity, specificity, and F-1 score (a measure of balance between precision and recall) were calculated. The best-performing model was described using the ratio of observed to expected outcomes, calibration slope, and calibration-in-the-large (CITL). Further, calculated probabilities were used to create a scoring system for embryos. A score of 1 equated to the worst prognosis and a score of 10 the best. Scoring thresholds were defined by ensuring there was an equal frequency of embryos in each score.

Results

Participants

A total of 3004 euploid embryos (37.4%) and 5023 aneuploid embryos (62.6%) were included in the final verified dataset; 120 (1.5%) mosaic embryos were excluded. Overall, 35 features (17 morphokinetic, 10 embryological, and 8 clinical variables) were extracted from electronic records. Table I describes participant characteristics as a proportion of euploid and aneuploid embryos. Figure 1 illustrates the number of embryos contributing to each dataset, included features and validation process. Supplementary Table SIII summarizes the variables included in each dataset and the proportion of missing data for each variable.

Univariable analyses

All morphokinetic variables were significantly delayed in aneuploid embryos apart from tPB2 and tPNa. The three morphokinetic variables with the largest median difference between euploid and aneuploid were tB (3 h, $P < 0.001$), tEB (2.5 h, $P < 0.001$), and tHB (1.93 h, $P < 0.001$) (Table I). Univariable analysis demonstrated that the likelihood of euploidy was significantly increased the more expanded the blastocyst and the better the trophectoderm grade. Furthermore, no association with ploidy status was reported for morula or cleavage stage fragmentation, morula grade, IVF or ICSI, sperm concentration, or progressive motility. The presence of multinucleation at the two-cell or four-cell stage was not associated with ploidy status of the blastocyst (Table I). Male ageing did not correlate with the percentage of euploid embryos when stratified for female age or in the regression analysis (Supplementary Fig. S1).

Modelling results

The multivariable LR model revealed that tM, tB, FSH dose, long or short protocol, and female age were significantly associated with ploidy status after adjustment (Supplementary Table SIV). The final euploidy predictors after backward elimination using LR were: t6, t7, tSC, tM, tB, FSH dose, female age, and stimulation protocol type (Supplementary Table SV). The model performance metrics of all 12 models are displayed in Table II. The sensitivity analyses using k-

Table I Univariable analysis of unadjusted associations between participant characteristics, embryological features, and ploidy status.

	Euploid	Aneuploid	*P*-value
Number of embryos (%)	3004 (37.4%)	5023 (62.6%)	
Mean oocyte provider age, years, (SD)*	36.5 (3.45)	38.7 (3.55)	<0.001
Oocyte provider age (years)**			
<35	740 (54.1%)	627 (45.9%)	<0.001
35–37	978 (49.3%)	1007 (50.7%)	
38–39	565 (36.5%)	981 (63.5%)	
40–42	529 (23.7%)	1704 (76%)	
≥43	61 (9.5%)	584 (90.5%)	
BMI, kg/m^2, median (IQR)*	23.5 (21.5–26.7)	23.5(21.5–26.8)	0.815
AMH (pmol/l), median (IQR)*	16.8 (9.7–27.2)	15.5 (8.8–25.3)	0.003
AFC, median (IQR)*	10 (0–18)	10 (0–19)	0.801
Number of oocytes retrieved, median (IQR)*	14 (10–19)	13 (9–18)	<0.001
Sperm provider age, years, mean (SD)*	38.7 (5.35)	40.1 (5.6)	<0.001
Protocol type**			<0.001
Short	2568 (85.5%)	4023 (80.1%)	
Long	436 (14.5%)	1000 (19.9%)	
Average FSH dose received**			<0.001
75–150 IU	766 (49.1%)	793 (50.9%)	
187.5–225 IU	664 (38.9%)	1044 (61.1%)	
262.5–337.5 IU	1126 (32.7%)	2320 (67.3%)	
≥450 IU	268 (29.0%)	657 (71.0%)	
Embryos created from donated oocyte*	143 (1.8%)	126 (1.6%)	0.87
Embryos created from donated sperm*	207 (2.6%)	413 (5%)	0.001
Genetic platform used for testing**			
aCGH	109 (3.6%)	258 (5%)	0.02
NGS	2895 (96.5%)	4765 (95%)	
Morphokinetics (median hours post insemination) and IQR*			
tPB2	3.16 (2.53–3.88)	3.10 (2.51–3.82)	0.52
tPNa	7.51 (6.24–9.3)	7.43 (6.21–9.28)	0.69
tPNf	23.2 (21.46–24.9)	23.38 (21.73–25.4)	<0.001
t2	25.71 (24.07–27.65)	25.95 (24.25–28.06)	<0.001
t3	36.54 (34.16–39.06)	36.89 (34.51–39.54)	<0.001
t4	37.7 (35.46–40.37)	38.18 (35.8–40.93)	<0.001
t5	49.28 (45.6–53)	49.71 (45.83–53.7)	<0.001
t6	51.4 (48.15–55.48)	51.98 (48.41–55.89)	<0.001
t7	53.44 (49.84–58.01)	54.1 (50.18–59.36)	<0.001
t8	56.53 (51.81–64.15)	57.2 (52.4–65.05)	<0.001
t9	69.8 (64.13–75.64)	70.55 (64.82–76.36)	<0.001
tSC	80.73 (75.1–86.62)	81.34 (75.5–87.52)	<0.001
tM	88.78 (83.07–94.83)	89.49 (83.75–95.72)	<0.001
tSB	97.28 (92.01–103.3)	99.02 (93.7–105.15)	<0.001
tB	105.02 (99.99–111.6)	108.02 (102.35–113.86)	<0.001
tEB	109.94 (104.19–114.83)	112.45 (106.92–118.59)	<0.001
tHB	111.04 (105.04–117.35)	112.97 (107.48–121.72)	<0.001
Morula fragmentation (number embryos, percentage)**			0.35
0–10%	516 (45.6%)	834 (43%)	
11–20%	394 (34.0%)	714 (36.9%)	

(continued)

Table I Continued

	Euploid	Aneuploid	P-value
21–50%	218 (18.8%)	345 (29.8%)	
>50%	30 (2.6%)	40 (2%)	
Cleavage stage fragmentation (number embryos, percentage)**			0.37
0–10%	2143 (74.9%)	3635 (76.2%)	
11–20%	540 (18.9%)	875 (18.3%)	
21–50%	117 (4.1%)	258 (5.4%)	
>50%	0	1 (0.02%)	
Grading at biopsy**			<0.001
Full blastocyst	308 (25.1%)	632 (31.1%)	
Expanded blastocyst	203 (16.5%)	365 (18%)	
Hatching blastocyst	717 (58.3%)	1035 (51%)	
ICM grade**			0.129
1	238 (18.5%)	361 (16.6%)	
2	899 (69.8%)	1511 (69.6%)	
3	151 (11.7%)	298 (13.7%)	
Trophectoderm grade**			0.016
1	171 (13.7%)	228 (10.9%)	
2	746 (59.6%)	1228 (58.8%)	
3	335 (26.7%)	632 (30.3%)	
Morula grade**			0.399
1	709 (48.7%)	1186 (47.3%)	
2	1321 (52.7%)	1321 (52.7%)	
ART method (n, %)**			0.875
IVF	1571 (52.3%)	2636 (52.5%)	
ICSI	1433 (47.7%)	2387 (47.5%)	
Sperm concentration (million per ml) (median, IQR)*	63 (35–94)	62 (35–93.4)	0.487
Sperm progressive motility (%) (median, IQR)*	33 (13–54)	32 (13–52)	0.234
Multinucleation**			
Non	345 (54%)	626 (57%)	0.08
Two-cell	238 (37%)	398 (36%)	
Four-cell	59 (9%)	71 (6%)	

*Mann–Whitney U test for statistical significance.
**Chi^2 test for significance.
Mosaic embryos excluded and therefore n = 8027 embryos compared. AFC: antral follicle count; aCGH: array comparative genomic hybridization; NGS: next-generation sequencing; tPNf: time to pronuclear fade; tn: time to n cells; tSC: time to start of compaction; tM: time to formation of Morula; tSB: time to start of blastulation; tB: time to formation of blastocyst; ICM: inner cell mass.

nearest neighbours for imputation and standardization rather than normalization led to no improvement in model performance metrics during cross-validation.

Considering the overall performance of the models using the F1 score and AUC, the best-performing model was LR built using Dataset 1 with aneuploidy as the target outcome. This resulted in a meta-analysed AUC of 0.71 (95% CI 0.67–0.73) and F1 score of 0.77 (Supplementary Fig. S2). For the target of euploidy, as per the primary analyses, the AUC remained unchanged but the F1 score dropped to 0.59. The difference in discriminating between euploid and aneuploid embryos can be realized by examining the proportions in each risk score (n = 696 in each score 1–10) (Fig. 2). All AI models had inferior performance to the more traditional statistical approach. The best-performing model had a meta-analysed ratio of expected to observed probabilities of 1.049 (95% CI 0.859–1.239), calibration slope of 1.011 (95% CI 0.826–1.195), and a CITL of −0.039 (95% CI −0.309 to 0.213).

The model was unable to discriminate between euploid and mosaic embryos, and the LR model would most likely classify mosaic embryos as euploid. The best-performing model, LR, had a different performance depending on oocyte provider age; in a younger cohort of <35-year-olds the AUC was 0.58, whereas in patients aged 35–38 and ≥39 years the AUC was 0.62 and 0.74, respectively. As part of the exploratory analysis, a model was created using only the elements from Gardner's classification. This 'morphological' model (n = 3198 embryos) had an AUC of 0.53 and accuracy of 52%. Adjusting

Table II Model performance metrics by algorithm and dataset used.

Algorithm	Logistic regression			Random forest classifier		
Performance metric	**Dataset 1 Target = euploid**	**Dataset 2 Target = euploid**	**Dataset 1 Target = aneuploid**	**Dataset 1 Target = euploid**	**Dataset 2 Target = euploid**	**Dataset 1 Target = aneuploid**
Sensitivity (%)	91	72	84	34	34	84
Specificity (%)	29	34	35	85	82	37
Positive predictive value (%)	43	39	68	57	54	69
Negative predictive value (%)	85	67	56	68	67	58
False positive rate (%)	71	66	65	15	18	63
False negative rate (%)	9	28	16	66	66	16
AUC (external)	0.71	0.69	0.71	0.63	0.63	0.68
Accuracy (%)	62	48	65	66	64	66
Precision (%)	43	40	68	56	67	69
F-1 score	0.68	0.51	0.77	0.42	0.74	0.76
Algorithm	**XGBoost**			**Deep learning**		
Performance metric	**Dataset 1 Target = euploid**	**Dataset 2 Target = euploid**	**Dataset 1 Target = aneuploid**	**Dataset 1 Target = euploid**	**Dataset 2 Target = euploid**	**Dataset 1 Target = aneuploid**
Sensitivity (%)	70	71	82	80	81	63
Specificity (%)	58	59	34	40	41	63
Positive predictive value (%)	50	52	67	69	69	50
Negative predictive value (%)	76	76	54	55	56	74
False positive rate	42	41	66	60	56	38
False negative rate	30	29	18	20	59	37
AUC (external)	0.55	0.61	0.63	0.56	0.58	0.63
Accuracy	62	63	64	65	66	63
Precision	49.9	52	67	69	69	50
F-1 score	0.58	0.68	0.74	0.74	0.74	0.56

AUC: area under the receiver operator curve; XGBoost: extreme gradient boosting.

morphokinetic timings to tPNf = 0 or using interval times worsened all model performance metrics. Excluding embryos analysed by aCGH did not improve model performance.

Discussion

The use of AI algorithms in reproductive medicine is becoming increasingly popular, however, a common limitation of published models is data insufficiency (Dimitriadis *et al.*, 2022). Concurrently, it is known that AI requires large amounts of data in order to provide reliable predictive performance (Curchoe *et al.*, 2020). With the advent of AI algorithms, we questioned whether modern AI techniques are superior in numerical data analysis for ploidy prediction in a large dataset of 8147 embryos. We present 12 models created using four algorithms on two datasets, each of which has been extensively validated. LR was the best performing algorithm, outperforming each of the AI models. Where data included more predictors but less embryos, the AUC worsened for the LR models but improved or was non-inferior for the AI models. This highlights the importance of large training datasets for AI modelling.

We have also reported additional findings, for example our data illustrated that male age is not associated with embryo ploidy status. Therefore, the previously published data reporting that male age is negatively associated with live birth is less likely to be secondary to embryo ploidy, considering the results presented here (Horta *et al.*, 2019). Furthermore, it was interesting that there was no improvement to AI model performance by performing feature selection and therefore all variables were included. It would be expected that by reducing the noise in the data it would lead to improved results. On the other hand, backward elimination was performed in the LR model, which did improve performance. The lack of improvement in AI models may be caused by underlying interactions between variables, a high prevalence of significant features or by a limited number of features being available. While we had a high number of predictors for a morphokinetic clinical model, feature selection may be more useful when there

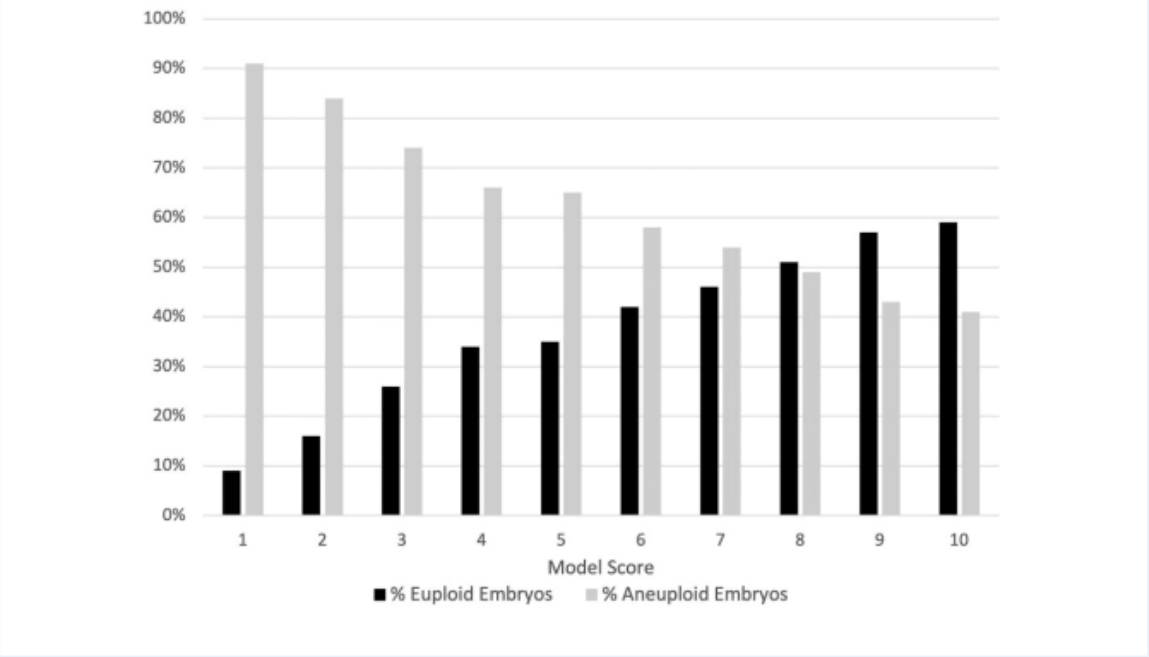


Figure 2. Bar chart showing the prevalence of euploid and aneuploid embryos for each risk score.

are many more variables in very large datasets. It is also interesting that model performance worsened by using tPNf as $t = 0$. This may, in part, be because if tPNf is the starting point, we exclude the important stages of fertilization, pronuclear formation and syngamy. Furthermore, the atypical juxtaposition and asynchronous breakdown of pronuclei are associated with abnormal first cleavage and impaired downstream development; this therefore could be reflected in the later timings, and the impact diluted if this phase is excluded (Ezoe *et al.*, 2022).

There is a drive to include blastocyst image analysis into AI embryo selection models. One notable model that does this devised a ranking algorithm using neural networks and reported an AUC of 0.74 (Chavez-Badiola *et al.* 2020). This study is limited by the size of the testing dataset ($n = 84$ embryos). Others report comparable results, however, also using small sample sizes (Huang *et al.* 2021). One study with impressive results had a large sample size ($n = 10\,638$), evaluating a DL model using time-lapse videos (Tran *et al.*, 2019). They reported limited performance metrics but an AUC of 0.93; however, prevalence of the target outcome (foetal heart) accounted for just 8% of the dataset. The AUC can, therefore, be misleading in heavily imbalanced datasets and will be the main reason why the AUC for ploidy prediction is typically lower than models predicting clinical outcomes.

One controversial issue associated with many types of AI algorithms is the so-called 'black box' models. These are models too complex for the human brain to understand or they are protected by intellectual property legislation (Rudin, 2019). This results in limited interpretability of algorithms, such as DL neural networks, and, importantly, confounders cannot be adjusted for. This can create challenges with the application of such models externally and the detection of errors within the models themselves (Ho *et al.*, 2020). It is therefore debatable whether we can ask embryologists to make decisions using a poorly understand model. This potentially creates issues with accountability for the decisions made and it is for these reasons that some authors argue that such modelling approaches should not be used in reproductive medicine (Afnan *et al.*, 2021). Furthermore, these approaches have as yet not been tested in a randomized controlled trial. Considering the results of this study and the limitations of a 'black box' approach, the arguments in favour of traditional statistical models are strengthened.

It is not solely euploidy that accounts for embryo quality and there are many other factors within an embryo which determine viability. In fact, morphokinetics has been associated with some of these other factors, for example mitochondrial content (Ho *et al.*, 2018). This may be where live birth models have an advantage and is an argument in favour of using this as the target outcome over ploidy status. These other elements related to embryo quality may lead to a greater difference between the morphokinetics of competent and incompetent embryos, potentially allowing for better discrimination. Therefore, the use of image analysis may become more important for ploidy prediction and integral to its future success. Importantly, whether morphokinetic models built for live prediction outperform ploidy models will be tested in an upcoming study. Furthermore, data imbalance may account for why many of the models we created coding for aneuploidy as the target variable performed better. It is for this reason we report the F1 score, which is less affected by such a phenomenon. This is

likely one of the reasons that our model performed better in older patients, whom have a higher proportion of aneuploid embryos, for which the model is better at defining. In other words, the best-performing model is sensitive but not specific for detecting aneuploidy.

This study provides evidence to support the use of time lapse. We have seen how morphokinetic parameters add valuable prognostic information, many of the which were ranked highly by our feature selection algorithm (Supplementary Fig. S3). In fact, if we were to include only morphokinetic predictors into the LR model, the AUC would be 0.61 and F1 score 0.72. Similarly, the 'morphological model', developed as part of an exploratory analysis, had an inferior AUC for ploidy prediction compared to the 'morphokinetics only' model with an accuracy of only 52% compared to 65%, respectively. This suggests that morphokinetic models are superior to embryo grading for ploidy prediction. Liu *et al.* (2022) developed a morphological model that had an AUC of 0.71, however, they also incorporated clinical characteristics, many of which are known to be the best predictors of ploidy status at the level of the woman. The controversies surrounding the use of a TLS may, in part, be because comparing different centres using time lapse is challenging; there are considerable differences in laboratory practices and the models which are being compared. Some centres use basic hierarchical or decision tree models, while others use complex machine learning methods. This potentially explains the results reported by Armstrong *et al.*, (2019). As embryo selection models become more complex with greater data acquisition, further randomized controlled trials are needed to test these newer embryo selection methods.

This study has highlighted the subtle differences in the development of aneuploid versus euploid embryos. These differences can be detected by a TLS; therefore, quality assurance of these annotations is essential as variations could potentially affect the fate of an embryo. A recent study at the participating centres included 59 embryologists who were required to fully annotate the same three embryos. It was concluded that there was very strong agreement using intraclass correlation coefficient (0.81–1.0) for all operators except one (Barrie, 2021). This supports the detection of these subtle differences; arguably making TLS superior to the day of blastocyst formation for embryo selection.

One of the strengths of our study is the size of the dataset, it may be for this reason that our results contradict those found by recent studies. For instance, one study compared multiple machine learning techniques, and the authors selected RFC as the best performing algorithm, however, the dataset used included only 539 embryos (De Gheselle *et al.*, 2022). Their conclusion that LR is inferior to AI algorithms is therefore not unsurprising as, for instance, when we performed a complete cases analysis using all extracted predictors ($n = 690$ embryos) the algorithm performed considerably worse (AUC 0.56). A second strength of this study is the robust validation processes implemented, using techniques such as internal–external validation to repeatedly validate the model. Many commercially available models lack published external validation studies and therefore critiquing performance is challenging. Where studies are available, they may use biased methods, such as validation on imputed datasets or 'apparent validation', whereby assessment of performance is derived from the development cohort, leading to model overestimation (Basile *et al.*, 2014; Chawla *et al.*, 2015; Del Carmen Nogales *et al.*, 2017; Desai *et al.*, 2018). Third, compared to known implantation data, ploidy data is much more balanced and therefore statistics such as AUC can be more reliably interpreted. Finally, the ability of our model to non-invasively detect aneuploid embryos with an acceptable degree of accuracy is a meaningful development for the use of morphokinetics.

One of the limitations of our study is the lack of image or video analysis; this analysis could have strengthened our model's performance. Second, this model has been shown to have a good degree of accuracy for selecting out aneuploid embryos, however, whether that translates into improved clinical outcomes remains to be determined. This is challenging to establish within this dataset as no aneuploid embryos will have been transferred. Therefore, a further prospective study will examine miscarriage and live birth rates across different model risk scores. Third, a challenge for prognostic models in reproductive medicine is to ultimately predict chance of live birth, therefore disregarding the endometrium could be seen as a limitation. That said, the aim of this model is prioritizing embryos with the highest chance of euploidy. While incorporating biomarkers of endometrial receptivity could give a more accurate live birth prediction at the level of the patient, it will have no benefit at the level of the embryo to improve selection. In fact, the addition of more clinical variables into prognostic models may mean that there is little variation in risk scores within a patient's cohort of embryos as the model may place more weight on such features. The existence of this phenomenon will also be examined in our next clinical study. Following this, the model's ability to rank embryos will be assessed, comparing to random allocation and senior embryologists. Finally, owing to limited data on direct or reverse cleavage we could not include this feature in any modelling; a systematic review both by the authors of this study and others has shown that this may be associated with ploidy status and therefore not including it could be a significant limitation (Liu *et al.*, 2014; Bamford *et al.*, 2022). That said, the prevalence of such abnormal cleavages has been reported to be as low as 4–26% and therefore it may be that including this variable impacts the predictions of only a small proportion of embryos (Barrie *et al.*, 2017; Ozbek *et al.*, 2021).

Conclusion

This study has highlighted the ability to harness morphokinetics and clinical data as a tool for ploidy prediction using one of the largest datasets known. Several machine learning algorithms were compared, the best performing was LR with an AUC of 0.71 and F1 score of 0.77 without the aid of blastocyst image analysis. An LR model including only embryo grading parameters was unable to discriminate between euploid and aneuploid embryos, whereas a model only incorporating morphokinetic predictors had an AUC of 0.61 for determining ploidy status. Blastocyst expansion and trophectoderm grade are the only morphological characteristics with prognostic signal for ploidy. Reduced sperm concentration, motility or male age are not associated with the chance of euploid blastocyst formation. This model may aid decision-making, particularly where PGT-A is not permitted or for prioritizing embryos for biopsy.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;**69**:245–247.

Tabachnick BG, Fidell LS, Ullman JB. *Using Multivariate Statistics*. Boston, MA: Pearson, 2007.

Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum Reprod* 2019;**34**:1011–1018.

Triguero I, García-Gil D, Maillo J, Luengo J, García S, Herrera F. Transforming big data into smart data: an insight on the use of the

Data availability

The data underlying this article cannot be shared publicly due to intellectual property rights. Aspects of the data will be shared on reasonable request to the corresponding author. Much of the data is provided within this article and Supplementary Material.

Acknowledgements

We would like to thank Lucinda Archer, Lecturer in Biostatistics at Keele University for her support in model development and validation. Further, we would like to thank the embryologists involved in data collection: Selina Young, Anna Brodie, Rafaela Iasonos, Gina Aldis, Audrey Watcher, and Sharon Corcoran. The first author would like to thank Dr Victoria Sephton (Care Group Medical Director), Professor Charles Kingsland (Care Chief Medical Officer), and Mr Philip Lowe for support and supervision during this PhD fellowship.

Authors' roles

T.B.: conception and design, analysis, and interpretation of the data, drafting the article, final approval of the version to be published. C.E.: statistical support for analysis and interpretation of the data, defined model building methodology, final approval of version to be published. S.M.: critical revision for important intellectual content, final approval of version to be published. R.S.: acquisition of data and compiling of datasets, final approval of version to be published. R.K.D.-S.: critical revision of important intellectual content, final approval of version to be published. A.B.: provided expert opinion in the revision of and final approval of version to be published. A.Ca.: provided expert opinion in the revision of and final approval of version to be published. A.Co.: chief investigator and final approval of version to be published.

Funding

No specific funding was sought for this study, university funds supported the first author throughout the study period paid by the University of Birmingham as part of his PhD grant.

Conflict of interest

A.Ca. is a minor shareholder of participating centres. Time lapse is a technology for which patients are charged extra at participating centres.

References

Afnan MAM, Liu Y, Conitzer V, Rudin C, Mishra A, Savulescu J, Afnan M. Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Hum Reprod Open* 2021;**2021**: hoab040.

Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database Syst Rev* 2019;**5**: CD011320.

Ayyadevara VK. Random forest. In: Ayyadevara VK (ed). *Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R*. Berkeley, CA: Apress, 2018, 105–116.

Bamford T, Barrie A, Montgomery S, Dhillon-Smith R, Campbell A, Easter C, Coomarasamy A. Morphological and morphokinetic associations with aneuploidy: a systematic review and meta-analysis. *Hum Reprod Update* 2022;**28**:656–686.

Barrie A. P165 Design, implementation and results of a group-wide, embryo morphokinetic annotation quality assurance scheme across ten fertility clinics. Fertility 2021 Barriers and breakthroughs 6–10th January 2021 Online. *Hum Fertil* 2021;**24**:46–69.

Barrie A, Homburg R, McDowell G, Brown J, Kingsland C, Troup S. Preliminary investigation of the prevalence and implantation potential of abnormal embryonic phenotypes assessed using time-lapse imaging. *Reprod Biomed Online* 2017;**34**:455–462.

Basile N, Nogales Mdel C, Bronet F, Florensa M, Riqueiros M, Rodrigo L, García-Velasco J, Meseguer M. Increasing the probability of selecting chromosomally normal embryos by time-lapse morphokinetics analysis. *Fertil Steril* 2014;**101**:699–704.

Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.

Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Hickman CF. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. *Reprod Biomed Online* 2013a;**26**: 477–485.

Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Thornton S. Retrospective analysis of outcomes after IVF using an aneuploidy risk model derived from time-lapse imaging without PGS. *Reprod Biomed Online* 2013b;**27**:140–146.

Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reprod Biomed Online* 2020;**41**: 585–593.

Chawla M, Fakih M, Shunnar A, Bayram A, Hellani A, Perumal V, Divakaran J, Budak E. Morphokinetic analysis of cleavage stage embryos and its relationship to aneuploidy in a retrospective time-lapse imaging study. *J Assist Reprod Genet* 2015;**32**:69–75.

Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health* 2020;**8**:e000262.

Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, Logullo P, Beam AL, Peng L, Van Calster B *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;**11**:e048008.

Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BJOG* 2015;**122**: 434–443.

Cornelisse S, Zagers M, Kostova E, Fleischer K, van Wely M, Mastenbroek S. Preimplantation genetic testing for aneuploidies (abnormal number of chromosomes) in in vitro fertilisation. *Cochrane Database Syst Rev* 2020;**9**:CD005291.

Curchoe CL, Malmsten J, Bormann C, Shafiee H, Flores-Saiffe Farias A, Mendizabal G, Chavez-Badiola A, Sigaras A, Alshubbar H, Chambost J *et al.* Predictive modeling in reproductive medicine:

where will the future of artificial intelligence research take us? *Fertil Steril* 2020;**114**:934–940.

Cutler A, Cutler DR, Stevens JR. Random forests. In: Zhang C, Ma Y (eds). *Ensemble Machine Learning: Methods and Applications*. Boston, MA: Springer US, 2012, 157–175.

De Gheselle S, Jacques C, Chambost J, Blank C, Declerck K, De Croo I, Hickman C, Tilleman K. Machine learning for prediction of euploidy in human embryos: in search of the best-performing model and predictive features. *Fertil Steril* 2022;**117**:738–746.

Del Carmen Nogales M, Bronet F, Basile N, Martinez EM, Linan A, Rodrigo L, Meseguer M. Type of chromosome abnormality affects embryo morphology dynamics. *Fertil Steril* 2017;**107**:229–235.e2.

Desai N, Goldberg JM, Austin C, Falcone T. Are cleavage anomalies, multinucleation, or specific cell cycle kinetics observed with time-lapse imaging predictive of embryo developmental capacity or ploidy? *Fertil Steril* 2018;**109**:665–674.

Dimitriadis I, Zaninovic N, Badiola AC, Bormann CL. Artificial intelligence in the embryology laboratory: a review. *Reprod Biomed Online* 2022;**44**:435–448.

Ezoe K, Coticchio G, Takenouchi H, Taoda S, Namerikawa S, Honda K, Miki T, Okimura T, Kobayashi T, Borini A *et al.* Spatiotemporal perturbations of pronuclear breakdown preceding syngamy affect early human embryo development: a retrospective observational study. *J Assist Reprod Genet* 2022;**39**:75–84.

Gardner DK. In vitro culture of human blastocyst. In: Mortimer D (ed). *Towards Reproductive Certainty: Infertility and Genetics Beyond 1999*. Carnforth: Parthenon Press, 1999,377–388.

Ho JR, Arrach N, Rhodes-Long K, Salem W, McGinnis LK, Chung K, Bendikson KA, Paulson RJ, Ahmady A. Blastulation timing is associated with differential mitochondrial content in euploid embryos. *J Assist Reprod Genet* 2018;**35**:711–720.

Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns (N Y)* 2020;**1**:100129.

Horta F, Vollenhoven B, Healey M, Busija L, Catt S, Temple-Smith P. Male ageing is negatively associated with the chance of live birth in IVF/ICSI cycles for idiopathic infertility. *Hum Reprod* 2019;**34**: 2523–2532.

Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can predict embryo ploidy status based on time-lapse data. *Reprod Biol Endocrinol* 2021;**19**:185.

Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. *Appl Artif Intell* 2019; **33**:913–933.

Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med Res Methodol* 2017;**17**:10.

Liu Y, Chapple V, Roberts P, Matson P. Prevalence, consequence, and significance of reverse cleavage by human embryos viewed with the use of the Embryoscope time-lapse video system. *Fertil Steril* 2014;**102**:1295–1300.e2.

Liu Y, Ong K, Korman I, Turner R, Shaker D, Zander-Fox D, Rombauts L. The effect of day 5 blastocyst assessment timing on live birth prediction and development of a prediction algorithm. *Reprod Biomed Online* 2022;**44**:609–616.

Mumusoglu S, Yarali I, Bozdag G, Ozdemir P, Polat M, Sokmensuer LK, Yarali H. Time-lapse morphokinetic assessment has low to moderate ability to predict euploidy when patient- and ovarian stimulation-related factors are taken into account with the use of clustered data analysis. *Fertil Steril* 2017;**107**:413–421.e4.

Munné S, Kaplan B, Frattarelli JL, Child T, Nakhuda G, Shamma FN, Silverberg K, Kalista T, Handyside AH, Katz-Jaffe M *et al.* Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. *Fertil Steril* 2019;**112**:1071–1079.e7.

Musil CM, Warner CB, Yobas PK, Jones SL. A comparison of imputation techniques for handling missing data. *West J Nurs Res* 2002; **24**:815–829.

Neal SA, Morin SJ, Franasiak JM, Goodman LR, Juneau CR, Forman EJ, Werner MD, Scott RT. Preimplantation genetic testing for aneuploidy is cost-effective, shortens treatment time, and reduces the risk of failed embryo transfer and clinical miscarriage. *Fertil Steril* 2018;**110**:896–904.

Ozbek IY, Mumusoglu S, Polat M, Bozdag G, Sokmensuer LK, Yarali H. Comparison of single euploid blastocyst transfer cycle outcome derived from embryos with normal or abnormal cleavage patterns. *Reprod Biomed Online* 2021;**42**:892–900.

Peng C-YJ, Harwell M, Liou S-M, Ehman LH. Advances in missing data methods and implications for educational research. *Real Data Anal* 2006;31–78.

Pirtea P, De Ziegler D, Tao X, Sun L, Zhan Y, Ayoubi JM, Seli E, Franasiak JM, Scott RT. Jr., Rate of true recurrent implantation failure is low: results of three successive frozen euploid single embryo transfers. *Fertil Steril* 2021;**115**:45–53.

Potdar K, Pardawala TS, Pai CD. A comparative study of categorical variable encoding techniques for neural network classifiers. *Int J Comput Appl* 2017;**175**:7–9.

Pribenszky C, Nilselid A-M, Montag M. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. *Reprod Biomed Online* 2017;**35**:511–520.

Putatunda S, Rama K. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. In: *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*. Shanghai, China: Crowne Plaza Shanghai, 2018, 6–10.

Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. *Stat Med* 2019;**38**:1276–1296.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;**1**:206–215.

Sanders KD, Silvestri G, Gordon T, Griffin DK. Analysis of IVF live birth outcomes with and without preimplantation genetic testing for aneuploidy (PGT-A): UK Human Fertilisation and Embryology Authority data collection 2016–2018. *J Assist Reprod Genet* 2021; **38**:3277–3285.

Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *J R Stat Soc Ser C Appl Stat* 1999;**48**: 313–329.

CHAPTER 4: ASSOCIATION BETWEEN A MORPHOKINETIC PLOIDY PREDICTION MODEL RISK SCORE AND MISCARRIAGE AND LIVE BIRTH: A MULTICENTRE COHORT STUDY

Thomas Bamford, Rachel Smith, Christina Easter, Rima Dhillon-Smith, Amy Barrie, Sue Montgomery, Alison Campbell, Arri Coomarasamy

Manuscript published in Fertility and Sterility, 10th June 2023

https://doi.org/10.1016/j.fertnstert.2023.06.006

My role in this publication is as follows: I collected and cleaned all the data used in this study. I performed all statistical analyses and wrote the whole of the manuscript.

Amendments: For this study only single embryo transfers from fresh IVF/ICSI cycles were included. There was more than one cycle included for some participants, controlled for using mixed effects.

COHORT STUDY

Association between a morphokinetic ploidy prediction model risk score and miscarriage and live birth: a multicentre cohort study

Thomas Bamford, M.B., Ch.B,[a,b] Rachel Smith, B.Sc.,[b] Christina Easter, M.Sc.,[a] Rima Dhillon-Smith, Ph.D.,[a] Amy Barrie, Ph,D.,[b] Sue Montgomery, Ph.D.,[b] Alison Campbell, Ph.D.,[b] and Arri Coomarasamy, M.D.[a]

[a] Tommy's National Centre for Miscarriage Research, Institute of Metabolism and Systems Research, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom; CARE Fertility Manchester, Manchester, Greater Manchester, United Kingdom; [b] CARE Fertility Headquarters, Nottingham, United kingdom

Objective: To determine whether the aneuploidy risk score from a morphokinetic ploidy prediction model, Predicting Euploidy for Embryos in Reproductive Medicine (PREFER), is associated with miscarriage and live birth outcomes.
Design: Multicentre cohort study.
Setting: Nine in vitro fertilization clinics in the United Kingdom.
Patients: Data were obtained from the treatment of patients from 2016–2019. A total of 3587 fresh single embryo transfers were included; preimplantation genetic testing for aneuploidy) cycles were excluded.
Intervention: PREFER is a model developed using 8,147 biopsied blastocyst specimens to predict ploidy status using morphokinetic and clinical biodata. A second model using only morphokinetic (MK) predictors was developed, P PREFER-MK. The models will categorize embryos into the following three risk score categories for aneuploidy: "high risk," "medium risk," and "low risk."
Main Outcome Measures: The primary outcomes are miscarriage and live birth. Secondary outcomes include biochemical clinical pregnancy per single embryo transfer.
Results: When applying PREFER, the miscarriage rates were 12%, 14%, and 22% in the "low risk," "moderate risk," and "high risk" categories, respectively. Those embryos deemed "high risk" had a significantly higher egg provider age compared with "low risk," and there was little variation in risk categories in patients of the same age. The trend in miscarriage rate was not seen when using PREFER-MK; however, there was an association with live birth, increasing from 38%–49% and 50% in the "high risk," "moderate risk," and "low risk" groups, respectively. An adjusted logistic regression analysis demonstrated that PREFER-MK was not associated with miscarriage when comparing "high risk" to "moderate risk" embryos (odds ratio [OR], 0.87; 95% confidence interval [CI], 0.63–1.63) or "high risk" to "low risk" embryos (OR, 1.07; 95% CI, 0.79–1.46). An embryo deemed "low risk" by PREFER-MK was significantly more likely to result in a live birth than those embryos graded "high risk" (OR, 1.95; 95% CI, 1.65–2.25).
Conclusion: The PREFER model's risk scores were significantly associated with live births and miscarriages. Importantly, this study also found that this model applied too much weight to clinical factors, such that it could no longer rank a patient's embryos effectively. Therefore, a model including only MKs would be preferred; this was similarly associated with live birth but not miscarriage. (Fertil Steril® 2023;■:■–■. ©2023 by American Society for Reproductive Medicine.)
Key Words: Morphokinetics, ploidy, model, artificial intelligence

Received February 24, 2023; revised May 31, 2023; accepted June 5, 2023.
Correspondence: Thomas Bamford, M.B., Ch.B, CARE Fertility Manchester, 108-112 Daisy Bank Rd, Manchester, Greater Manchester, M14 5QH, United Kingdom (E-mail: tom.bamford@carefertility.com).

Fertility and Sterility® Vol. ■, No. ■, ■ 2023 0015-0282/$36.00

https://doi.org/10.1016/j.fertnstert.2023.06.006

The last decade of assisted reproduction has seen a phenomenal change in the tools that an embryologist has at their disposal for embryo selection. The fundamentals of morphological assessment that were described in the 1990s remain an important marker of implantation potential (1). That said, with increasing access to "big data," machine learning algorithms can be used to consider a huge number of variables to assist with embryo selection (2, 3). Morphokinetic (MK) data derived from time-lapse systems (TLS) has therefore provided a wealth of data, postulated to be a marker for embryo quality. The annotations of an embryo's developmental milestones (i.e., MKs) can be compared with outcome variables, such as live birth or ploidy status. The feasibility of predicting the latter using MKs has been demonstrated in the first publication in this series (4). Prior research identified that aneuploid embryos are significantly delayed in a variety of MK variables, but most markedly around the time of blastulation. This data, combined with further clinical and embryological predictors, could be used to prioritize embryos for transfer, thus stratifying the risk of aneuploidy. The justifications for focusing on ploidy status over live birth have been outlined in the second article of this series (5). In brief, ploidy status is less vulnerable to confounders that exist in live birth data, such as endometrial or uterine factors. Second, the use of preimplantation genetic testing for aneuploidy (PGT-A) is increasing internationally, despite controversies; therefore, finding accessible, noninvasive alternatives remains an important venture (6). Finally, although there is a lack of consensus regarding PGT-A, studies have demonstrated higher ongoing pregnancy rates per embryo transferred, reduced time to pregnancy, and a reduced miscarriage rate, particularly in patients aged >35 years (7–10). Deselecting these embryos, therefore, seems advantageous and important for the patient.

There are several published models, each with its own limitations in terms of development and validation. One common limitation is the lack of data used to build the model (5–13). Existing models report an accuracy of between 60% and 70%, using a combination of TLS incubation data, other embryological or clinical data, and blastocyst image analysis using artificial intelligence (11–19). A meta-analysis and systematic review published in 2017 reported not only an improved chance of live birth but also significantly reduced early pregnancy losses when using a TLS (20). A Cochrane review published 2 years later included four additional studies. It was similarly reported that using a TLS for embryo selection may reduce miscarriage rates, although this finding was uncertain because of very low quality evidence (21). Although the Cochrane review failed to conclude that a TLS is associated with improved live birth rates, it must be noted that assessing this technology using meta-analysis is challenging (21). One may argue that including different TLS algorithms in the intervention cohort is akin to comparing the control group to a heterogeneous group of different treatment modalities. That said, this still provides some reassuring data for its use and, importantly, introduces an interesting concept for its application as a counseling tool, potentially providing the patient with an individualized estimate of miscarriage risk.

Twelve MK models were recently built using four different machine learning algorithms. One was selected as the best-performing algorithm for predicting ploidy after validation in a cohort of 8,147 biopsied blastocyst specimens (5). This model will be referred to as Predicting Euploidy for Embryos in Reproductive Medicine (PREFER). This study therefore aimed to determine whether a MK ploidy prediction model can stratify embryos for miscarriage risk whereas predicting live birth. The inclusion of age is debated commonly in artificial intelligence circles because it is well known that age in particular is the best predictor of success in in vitro fertilization (IVF) (22). There is therefore a risk that an algorithm applies too much weight to this covariate, such that there is very little variance or ranking within one patient's embryos. This study investigated this ongoing debate by comparing a model that includes only MK predictors (PREFER-MK) with another that also includes clinical parameters (PREFER).

MATERIALS AND METHODS

Data for this cohort study was obtained from the treatment of patients from 2016–2019 at nine IVF clinics in the United Kingdom. There were no major changes in laboratory procedures during this study period. The participating centers are part of a fertility group providing both private and state-funded treatments, each conforming to the same laboratory practices. This study has received institutional review board and ethical approval. The reporting of findings is in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology guidelines (23).

Model Development and Validation

The model development and validation have been described in full elsewhere in a study designed to select the best algorithm (5). In brief, a sample of 8,147 biopsied blastocysts was used. The best-performing algorithm was mixed-effects logistic regression using the patient identifier as the random intercept, allowing for within-cluster variability. A total of 22 predictors were selected for inclusion in the model using backward elimination with a P value of $<.157$. The final model, PREFER, included time to 6 cells (t6), 7 cells (t7), start of compaction (tSC), formation of morula (tM), formation of blastocyst (tB), follicle-stimulating hormone (FSH) dose, egg provider age, and long or short protocol. Considering the overall performance of the models assessed using internal and external cross validation, PREFER had a meta-analyzed area under the curve (AUC) of 0.71 (95% confidence interval [CI], 0.67–0.73) and an F1 score of 0.77. A model developed using only MK predictors had an AUC of 0.61 and an F1 score of 0.72 (5).

The models were applied to place embryos into three risk categories; thresholds were created on the dataset of 8,147 embryos by ensuring that there was an equal frequency of predicted probabilities generated by the algorithm in each risk score. To test the trend between each model's predicted risk of aneuploidy and clinical outcomes, six thresholds were created for graphical analysis only.

TABLE 1

Participant characteristics for PREFER and PREFER-MK model risk scores.

	PREFER				PREFER-MK (Morphokinetics only)			
Covariate	**High risk aneuploid (n = 1,195)**	**Moderate risk aneuploid (n = 1,196)**	**Low risk aneuploid (n = 1,196)**	***P* value**	**High risk aneuploid (n = 1,195)**	**Moderate risk aneuploid (n = 1,196)**	**Low risk aneuploid (n = 1,196)**	***P* value**
Egg provider age (n, SD)	37 (2.38)	34 (1.79)	29 (3.1)	<.001	33 (4.3)	33 (4.5)	32.8 (4.3)	.2
Own eggs	1151 (40%)	967 (33.8%)	746 (26%)	<.001	1002 (35%)	937 (33%)	925 (32%)	<.001
Donor eggs	44 (6%)	229 (31.7%)	450 (62%)		193 (27%)	259 (36%)	271 (37%)	
Eggs retrieved	10.1 (4.89)	10.9 (5.42)	11.6 (5.41)	<.001	10.5 (5.1)	10.8 (5.2)	11.3 (5.4)	.06
Eggs fertilized (n, SD)	6.2 (3.18)	6.9 (3.6)	7.3 (3.7)	<.001	6.5 (3.36)	6.7 (3.4)	7.2 (3.7)	<.001
FSH dose				<.001				<.001
75–150 IU	231 (13.9%)	554 (33.4%)	874 (52.3%)		494 (29.8%)	562 (34%)	603 (63%)	
187.5–225 IU	330 (37.8%)	352 (40.3%)	192 (21.9%)		305 (34.9%)	293 (34%)	276 (32%)	
262.5–337.5 IU	364 (55.2%)	194 (29.4%)	101 (15.3%)		296 (36.7%)	260 (32%)	250 (31%)	
≥450 IU	169(68%)	64 (25.8%)	15 (6.1%)		100 (40.3%)	81 (33%)	67 (27%)	
Long protocol	602 (33%)	609 (33%)	595 (33%)	.849	591 (33%)	598 (34%)	617 (34%)	.556
Short protocol	593 (33%)	587 (33%)	601 (33%)		604 (34%)	598 (33%)	579 (33%)	
IVF	385 (32.6%)	414 (35%)	380 (32%)	.282	326 (28%)	418 (35%)	435 (37%)	<.001
ICSI	810 (33.6%)	782 (32.5%)	816 (33.9%)		869 (36%)	778 (32%)	761 (32%)	
BMI (n, SD)	25.12 (4.3)	24.9 (4.3)	24.3 (4.2)	.615	25 (3.9)	24.9 (3.6)	25 (3.8)	.725
AFC (median, IQR)	11 (7–19)	13 (6–22)	12 (2–24)	.003	13 (8–20)	12 (6–20)	13 (6–21)	.54
Endometrial thickness (mean, SD)	11.1 (2.3)	11.1 (2.3)	10.5 (2.2)	<.001	10.7 (2.71)	10.6 (2.9)	10.5 (3.2)	.26
Ethnicity				.009				.201
White British	891 (32%)	936 (33.7%)	954 (34.3%)		943 (33%)	959 (33%)	964 (34%)	
White Irish	26 (29%)	35 (39.7%)	27 (30.7%)		44 (29%)	43 (29%)	60 (41%)	
White and Black	25 (44%)	13 (22.8%)	19 (33.3%)		29 (33%)	28 (32%)	30 (34%)	
White and Asian	5 (35%)	5 (35.7%)	4 (28.5%)		8 (33%)	8 (33%)	6 (19%)	
East Asian	20 (48%)	12 (29.3%)	9 (22%)		6 (30%)	10 (50%)	4 (20%)	
South Asian	116 (36%)	112 (34.5%)	97 (29.8)		116 (36%)	108 (33%)	101 (31%)	
Black African or Caribbean	12 (35%)	4 (11.8%)	18 (53%)		9 (26%)	12 (35%)	13 (38%)	
Other	100 (40.5%)	79 (32%)	68 (27%)		40 (47%)	28 (33%)	18 (25%)	
Nulliparous	660 (31.7%)	689 (33.1%)	732 (35.2%)	.012	700 (34%)	685 (33%)	696 (33%)	.802
Parous	535 (35.5%)	507 (33.7%)	464 (30.8%)		495 (33%)	511 (34%)	500 (33%)	
History of ≥2 miscarriages	211 (35%)	191 (31.8%)	198 (33%)	.533	192 (32%)	212 (35%)	196 (33%)	.515
Use of sperm donor	130 (32.9%)	143 (36.2%)	122 (30.9%)	.384	104 (26%)	137 (35%)	154 (39%)	.004
Use of frozen sperm	209 (32.5%)	243 (37.8%)	191 (29.7%)	0.019	199 (31%)	219 (34%)	225 (35%)	.354
SSR sperm	10 (41.7%)	9 (37.5%)	5 (20.8%)	0.414	12 (50%)	6 (25%)	6 (25%)	.220
Blastocyst expansion				<.001				<.001
Very early	19 (51.3%)	7 (18.9%)	11 (29.7%)		16 (43%)	13 (35%)	8 (22%)	
Full	288 (43.1%)	213 (31.9%)	167 (25%)		316 (47%)	211 (32%)	141 (21%)	
Expanded	698 (30.7%)	757 (33.2%)	820 (36%)		725 (32%)	791 (35%)	759 (33%)	
Hatching	190 (31.3%)	219 (36.1%)	198 (32.6%)		138 (23%)	181 (30%)	288 (47%)	
ICM grade				.21				.09
A	155 (29.3%)	182 (34.5%)	191 (36.1%)		154 (29%)	172 (33%)	202 (38%)	
B	834 (33.6%)	820 (33.1%)	825 (33.2%)		844 (34%)	830 (33%)	805 (32%)	
C	206 (35.5%)	194 (33.4%)	180 (31%)		197 (34%)	194 (33%)	189 (33%)	

Bamford. Morphokinetic ploidy model validation. Fertil Steril 2023.

TABLE 1

Continued.

Covariate	PREFER				PREFER-MK (Morphokinetics only)			
	High risk aneuploid (n = 1,195)	Moderate risk aneuploid (n = 1,196)	Low risk aneuploid (n = 1,196)	*P* value	High risk aneuploid (n = 1,195)	Moderate risk aneuploid (n = 1,196)	Low risk aneuploid (n = 1,196)	*P* value
Trophectoderm grade				<.001				.002
A	61 (29.3%)	64 (30.8%)	83 (39.9%)		49 (24%)	70 (34%)	89 (43%)	
B	997 (33.3%)	1019 (34%)	974 (32.6%)		1014 (34%)	1013 (34%)	963 (32%)	
C	137 (35.2%)	113 (29.1%)	139 (35.7%)		132 (34%)	113 (29%)	144 (37%)	

Abbreviations: AFC = antral follicle count; BMI = body mass index; FSH = follicle-stimulating hormone; ICSI = intracytoplasmic sperm injection; IVF = in vitro fertilization; PREFER = Predicting Euploidy for Embryos in Reproductive Medicine; PREFER-MK = Predicting Euploidy for Embryos in Reproductive Medicine-morphokinetic.

Bamford Morphokinetic ploidy model validation. Fertil Steril 2023.

Participants

Patients were included in this study when they underwent an IVF or intracytoplasmic sperm injection (ICSI) cycle during the study period that included TLS incubation. Preimplantation genetic testing for aneuploidy cycles was excluded. Donor oocytes and autologous cycles were included; the age of the donor was included as well as the age of the egg provider. Double embryo transfers were excluded. Only fresh embryo transfers were included because the MKs could be applied to PREFER or PREFER-MK linking them to the clinical outcomes. All embryos were cultured in a TLS (EmbryoScope; Vitrolife, Sweden) at 37°C, 6%–6.5% CO_2, 5% O_2, and 88.5%–89% N_2. Morphokinetic parameters were annotated manually by an embryologist trained in-house, following strict annotation procedures validated elsewhere (24). Every 10 minutes, microscopic images were acquired of the embryos through seven multifocal planes. The wells of the Embryoscope slide (Vitrolife, Sweden) were filled with culture medium (Global Total LP; Lautrupparken, Denmark, Cooper Surgical) and covered with mineral oil (LifeGuard Oil; Cooper Surgical). All blastocysts were finally assessed for suitability for transfer on day 5 or 6. Embryo grading was performed according to Gardner classification (1).

Outcomes

The primary outcomes were miscarriage rate, defined as pregnancy loss up to 22 weeks after evidence of a clinical pregnancy (presence of an intrauterine gestational sac on an ultrasound scan at 6 weeks), and live birth. Secondary outcomes included biochemical pregnancy rate (defined as a positive pregnancy test) and clinical pregnancy per single embryo transfer. Each outcome was compared between each model's risk score as follows: "high risk aneuploid," "moderate risk aneuploid," and "low risk aneuploid."

Sample size

To calculate the sample size, the primary outcome of miscarriage was used because this required the largest sample size. This was determined by assessing three levels of risk for aneuploidy. An existing logistic regression MK model built to predict live birth was used to estimate prevalence by categorizing embryos into three groups of worst to best prognosis, in line with the PREFER model. Although euploidy and live birth are different outcomes, they are related, and therefore this was deemed the most reliable way to estimate the proportion of embryos transferred in each risk score. It was calculated that 9% "high risk aneuploid," 27% "moderate risk aneuploid," and 64% "low risk aneuploid" embryos would be transferred. Using the live birth model, the miscarriage rate for each risk score was predicted at 30%, 15%, and 10%, respectively. Given this information, the comparison between moderate risk and low risk embryos yielded the largest sample size required to achieve at least 90% power with an alpha value of 0.017, using the Bonferroni correction for multiple testing. This requires a total sample size of 3,027 embryos.

TABLE 2

Unadjusted clinical outcomes for each PREFER and PREFER-MK score.

	PREFER			PREFER-MK		
Outcome	High risk aneuploid (n = 1,195)	Moderate risk aneuploid (n = 1,196)	Low risk aneuploid (n = 1,196)	High risk aneuploid (n = 1,195)	Moderate risk aneuploid (n = 1,196)	Low risk aneuploid (n = 1,196)
Miscarriage rate	114/530 (22%)	100/700 (14%)	93/737 (12%)	90/549 (16%)	96/688 (14%)	121/730 (17%)
Biochemical pregnancy rate	607/1,195 (60%)	778/1,196 (65%)	810/1,196 (68%)	640/1,195 (54%)	752/1196 (63%)	803/1,196 (67%)
Clinical pregnancy rate	530/1,195 (44%)	700/1,196 (58%)	737/1,196 (61%)	549/1,195 (46%)	688/1,196 (58%)	730/1,196 (61%)
Live birth rate	409/1,195 (34%)	596/1,196 (49%)	638/1,196 (53%)	455/1,195 (38%)	589/1196 (49%)	599/1196 (50%)

Abbreviations: PREFER = Predicting Euploidy for Embryos in Reproductive Medicine, PREFER-MK = Predicting Euploidy for Embryos in Reproductive Medicine-morphokinetic.

Bamford. Morphokinetic ploidy model validation. Fertil Steril 2023.

Statistical Analyses

The relationship between the outcomes and the three risk levels determined using PREFER and PREFER-MK was assessed using multivariable logistic regression. The PREFER-MK score was adjusted for egg provider age, FSH dose, short or long protocol, body mass index, antral follicle count, endometrial thickness, ethnicity, parity, recurrent miscarriage history (three or more), use of ICSI, number of eggs retrieved and number fertilized, donor eggs, frozen thawed sperm, and surgically retrieved sperm. The PREFER score was already adjusted for age, long or short protocol, and FSH doses; therefore, these adjustments were excluded for this model. A sensitivity analysis was also conducted, adjusting the models for embryo quality (expansion, trophectoderm, and inner cell mass grade). This was excluded from the primary analyses because these models are a marker of embryo quality; therefore, this may represent any association between the risk scores and outcomes inaccurately. A second sensitivity analysis included only ICSI embryos to determine whether the model predicted clinical outcomes better in this cohort of embryos. An exploratory analysis tested whether a MK model could offer anything over and above age alone as a prognostic factor for live birth and miscarriage. This was tested by creating a model using only egg provider age as a predictor in an "age model" and comparing its performance to PREFER-MK and PREFER using the same methods as above. The influence of age within the models was explored further by examining the percentage of embryos given each of the six risk scores, as determined using PREFER and PREFER-MK, stratified for woman's age.

RESULTS

In the study period, there were 4,261 fresh embryo transfers meeting eligibility criteria. After including only single embryo transfers, a total of 3,587 blastocysts remained from the treatment of 3,405 patients. There was <5% missing data because all variables were recorded routinely as part of the clinic workflow during this study period. These data were therefore presumed to be missing at random, and no imputation was required (25, 26). Table 1 provides an overview of participant characteristics for each of the three aneuploidy risk scores when applying the PREFER or MKs only model (PREFER-MK). An embryo determined using the PREFER model as having "high risk aneuploidy" was associated with a 22% miscarriage rate, decreasing to 14% and 12% in the "moderate risk aneuploid" and "low risk aneuploid" groups, respectively (Table 2). For comparison, the participating centers' live birth rate and miscarriage rate for a single euploid embryo transfer are 50% and 15%, respectively. This trend in miscarriage rate was no longer seen when using PREFER-MK; however, similar to the PREFER model, the live birth rate increased from 38% to 49% and 50% in the "high," "moderate risk," and "low risk aneuploid" groups, respectively. A similar finding was observed in the logistic regression analyses; after adjustment of the aforementioned covariates, a PREFER "low risk aneuploid" embryo had 49% decreased odds of miscarriage than a "high risk aneuploid" embryo (odds ratio [OR], 0.51; 95% CI, 0.63–0.72; $P<.001$). In contrast, the MKs only model was not associated with miscarriage when comparing the "high risk" to the "moderate risk" (OR, 0.87; 95% CI, 0.63–1.63; $P=.39$) or "high" to "low risk aneuploid" embryos (OR, 1.07; 95% CI, 0.79–1.46; $P=.63$). Biochemical and clinical pregnancy rates all increased with a reduced likelihood of aneuploidy, as determined using both PREFER and PREFER-MK (Table 3). An embryo deemed "low risk aneuploid" using PREFER was over two times more likely to result in a live birth than those embryos graded "high risk aneuploid" (OR, 2.16; 95% CI, 1.79–2.62; $P<.001$). When applying the PREFER-MK, the association with live birth remained similar (OR, 1.95; 95% CI, 1.65–2.25; $P<.001$). Figure 1 illustrates that both PREFER and PREFER-MK perform similarly to a model using age as the sole predictor for live birth rate and miscarriage. Notably, those embryos deemed "high risk" using PREFER had a significantly higher age compared with "moderate" and "low risk"; this did not occur using PREFER-MK. Supplemental Table 1 (available online) illustrates how a model, including covariates such as age, results in little variance in risk scores within age categories. When only MKs are used, there is a spread of aneuploidy risk scores across each age category.

A sensitivity analysis revealed minimal diminishment of the effect size for all clinical outcomes after adjusting for embryo grading (Supplemental Table 2, available online). For instance, using the PREFER model, miscarriage remained more likely in the "high risk aneuploid" group compared

TABLE 3

Adjusted* analyses comparing each model's odds of miscarriage and live birth in the "high risk aneuploid" group, compared with "moderate" and "low risk aneuploid."

	PREFER model		Morphokinetics-only model		Age-only model	
Outcome	High risk vs. moderate risk (adjusted OR; 95% CI; *P* value)	High risk vs. low risk (adjusted OR; 95% CI; *P* value)	High risk vs. moderate risk (adjusted OR; 95% CI; *P* value)	High risk vs. Low risk (adjusted OR; 95% CI; *P* value)	High risk vs. moderate risk (adjusted OR; 95% CI; *P* value)	High risk vs. low risk (adjusted OR; 95% CI; *P* value)
Miscarriage rate	0.62; 95% CI, 0.45–0.85; $P = .003$	0.51; 95% CI, 0.36–0.72; $P < .001$	0.87; 95% CI, 0.63–1.63; $P = .39$	1.07; 95% CI, 0.79–1.46; $P = .63$	0.47; 95% CI, 0.34–0.66; $P < .001$	0.48; 95% CI, 0.33–0.66; $P < .001$
Live birth rate	1.84; 95% CI, 1.54–2.19; $P < .001$	2.16; 95% CI, 1.79–2.62; $P < .001$)	1.52; 95% CI, 1.29–1.8; $P < .001$	1.75; 95% CI, 1.38–2.23; $P < .001$	1.84; 95% CI, 1.54–2.22; $P < .001$	1.95; 95% CI, 1.65–2.25; $P < .001$

Abbreviations: BMI = body mass index, CI = confidence interval, FSH = follicle-stimulating hormone, ICSI = intracytoplasmic sperm injection, PREFER = Predicting Euploidy for Embryos in Reproductive Medicine, PREFER-MK = Predicting Euploidy for Embryos in Reproductive Medicine-morphokinetic.

* The PREFER-MK score was adjusted for egg provider age, FSH dose, short or long protocol, BMI, antral follicle count, endometrial thickness, ethnicity, parity, recurrent miscarriage history (three or more), use of ICSI, number of eggs retrieved and fertilized, donor eggs, frozen thawed sperm, and surgically retrieved sperm. The PREFER score was already adjusted for age, long or short protocol, and FSH doses; therefore, these adjustments were excluded for this model.

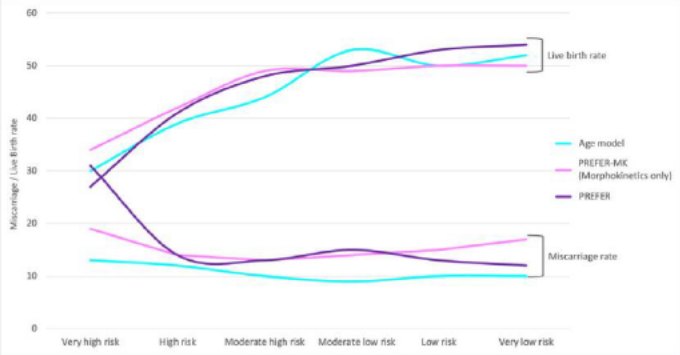
Bamford. Morphokinetic ploidy model validation. Fertil Steril 2023.

with the "moderate risk aneuploid" (OR, 0.63; 95% CI, 0.46–0.88; $P = .005$) and even more so when comparing the "low risk aneuploid" to the "high risk aneuploid" cohort (OR, 0.53; 95% CI, 0.37–0.75). As in the main analysis, PREFER-MK was not associated with miscarriage after adjusting for embryo quality but was associated with live birth (OR, 1.54; 95% CI, 1.25–1.72; $P < .001$). A second sensitivity analysis separated IVF and ICSI embryos, and the model scores were associated with similar significant improvements in pregnancy outcomes for both cohorts (Supplementary Table 3, available online). The PREFER score was not associated with miscarriage risk for IVF embryos when comparing the "high risk aneuploid" to the "moderate risk aneuploid" or "low risk aneuploid" scores, however, the effect size was more profound for ICSI embryos (OR, 0.97; 95% CI, 0.57–1.65; $P = .91$ and OR, 0.49; 95% CI, 0.31–0.72; $P < .001$ for IVF and ICSI, respectively) (Supplementary Table 3). When applying the MKs only model, there was no association with miscarriage when separating IVF and ICSI embryos (OR, 0.81; 95% CI, 0.44–1.47; $P = .49$ and OR, 0.91; 95% CI, 0.62–1.32; $P = .62$, respectively). PREFER-MK selected a different embryo for morphological selection 40% of the time.

DISCUSSION

These results provide evidence that a MK ploidy prediction model's risk score can be used to predict live birth but is not associated with miscarriage risk when only MKs are used. It has been demonstrated how age alone is significantly associated with live birth and miscarriage risk; indeed, many MK models would struggle to beat this predictor. It would appear at first that a MK model offers very little over and above age. However, it must be considered that age is a prognostic factor operating at the level of the woman, whereas MKs operate at the level of the embryo. When choosing the best embryo for a given patient, age is not going to be helpful. This concept has further been demonstrated by the lack of spread of risk scores or probabilities across patients of similar ages. It would therefore be prudent to use a model that does not account for age to allow for better ranking of embryos, i.e., the MKs only model, PREFER-MK. The association between both models' scores and all clinical outcomes when separating IVF and ICSI embryos remained significant, apart from the miscarriage rate for PREFER. It must be noted that this study was not powered for these analyses, and the miscarriage event rate for IVF embryos is lower than that for ICSI. The effect sizes in ICSI embryos were also slightly smaller, possibly because of an indirect effect of semen quality on live birth (27, 28). Furthermore, the sensitivity analysis demonstrated that the odds of live birth remain similar despite adjustments for embryo quality. It has also previously been found that a model incorporating only embryo grading was unable to discriminate aneuploid embryos (AUC = 0.52) (5). This is particularly interesting given that a different embryo is selected using the MK model over morphological selection 40% of the time. These considerations may suggest that the PREFER-MK score provides additional prognostic information over and above morphological grading. Finally, it has been reported that both models work similarly for live birth

FIGURE 1



Graph comparing the association between live birth rate and miscarriage for increasing the risk of aneuploidy, as predicted using PREFER, PREFER-MK, and the age-only models. Abbreviations: PREFER = Predicting Euploidy for Embryos in Reproductive Medicine, PREFER-MK = Predicting Euploidy for Embryos in Reproductive Medicine-morphokinetic.

Bamford. Morphokinetic ploidy model validation. Fertil Steril 2023.

prediction in IVF and ICSI embryos; therefore, the challenge of defining to in IVF embryos has not had a negative impact on model performance.

Although this study has shown that miscarriage risk is not associated with a MK model, it has clearly demonstrated strengths in live birth prediction. As such, this model could potentially be used as a counseling tool for patients with only poor prognosis embryos available to consider PGT-A. A MK model's performance will always be inferior to PGT-A because of the overlap between the timings of aneuploid and euploid embryos (4). For instance, the model has been shown previously to be better at discriminating aneuploid embryos, possibly accounting for the negligible difference observed in clinical outcomes between "low risk" and "very low risk" embryos (Fig. 1) (5).

There are potential explanations for why this study has demonstrated no association between miscarriage and MK risk scores in contrast to other studies (20, 21). First, this study was powered by the expected prevalence of miscarriage risk in embryos transferred. It is well documented that aneuploidy accounts for approximately 50% of miscarriages (29, 30). Although the primary outcome was miscarriage, in essence, the model is attempting to detect miscarriage of a chromosomally abnormal embryo or fetus, a subset of these miscarriages. Possibly, therefore, the sample size, whereas large, needed to be bigger to detect this difference. Additionally, it has been illustrated how PREFER-MK is sensitive (69%) but not specific for detecting aneuploidy (45%); therefore, it can be more certain of a negative outcome in the higher-risk categories but less certain in the lower risk categories. Finally, as highlighted by Armstrong et al. (21), although there is some signal that MK may be associated with preventing miscarriage, this evidence is of very low quality. Therefore, this study does provide evidence to contradict these findings.

Instrumenting MK models to predict ploidy status has had varying degrees of success (31–36). This is particularly the case in many of the randomized controlled trials reported to date. These studies found no difference in early pregnancy loss with the use of TLS models, yet they had a mean event rate of 4 (37–40). In other adequately powered trials, there have been significant differences, with the TLS group having a lower miscarriage rate and a higher live birth rate (41, 42). In contrast, one recent randomized trial reported no difference in miscarriage rate between TLS and morphological selection (relative risk [RR], 1.146; 95% CI, 0.788–1.668; n = 776) (43). Notably, this trial was underpowered; 911 additional participants were required for it to be adequately powered. Furthermore, the predictive performance of one model cannot be generalized to all MK models, particularly because MK has been found to be affected by local patient, laboratory, and clinical factors (44–50). It is therefore reasonable to suggest that models should be robustly validated in-house and, ideally, their effect on clinical outcomes should be prospectively tested.

A limitation of many available ploidy prediction models is sample size; the overriding strength of this study is the size of the dataset used to validate the effects of the model. Two datasets have been used; the first used 8,147 embryos to illustrate model performance for predicting ploidy status, whereas this study included 3,587 embryos to determine the clinical use of PREFER and PREFER-MK. The validation procedures presented are robust, encompassing internal and external cross validation and testing clinical applicability externally. There are further limitations to this study than those already discussed, primarily the use of only single embryo transfers, which can lead to an element of bias as those embryos transferred are likely to be of better quality compared with a cohort of double embryo transfers. That said, embryo quality has been adjusted for in the sensitivity analysis, and only a small proportion of double embryo transfers were excluded (n = 674). There are also limitations with retrospective study design; although many confounding variables have been accounted for, it remains impossible to eliminate such sources of bias completely. A further limitation is the exclusion of frozen embryo transfers; this eliminates a large proportion of data, which is potentially a cohort of slightly poorer-quality embryos. However, this was considered when powering this study on the three risk scores, ensuring enough embryos were transferred in the poor prognosis group to give reliable results. In addition, the clinical outcomes of participating centers remain similar between frozen and fresh transfers. The strength of MK models should ideally be tested using their ranking ability; previously, statistics such as sensitivity or accuracy have been reported. These are artificial in that this model will never be used to give a diagnosis. Considering this, a further study is planned to determine whether PREFER-MK is superior in ranking euploid embryos first compared with a model built on live birth and senior embryologists.

CONCLUSION

The PREFER model's risk scores were significantly associated with live births and miscarriages. Importantly, however, this study has demonstrated that age and clinical predictors can significantly influence a model such that it can no longer

competently rank embryos. To that end, a "MK only" ploidy model is the preferred approach, PREFER-MK. This model built using 8,147 biopsied embryos has been demonstrated to determine live birth but not miscarriage accurately in a dataset of 3,587 embryos. With a decreasing risk of aneuploidy, as determined using PREFER-MK, the live birth and clinical and biochemical pregnancy rates all increased. The true merit of a MK model is its ability to rank the most competent embryos first; this will now be tested in a study examining the ability of PREFER-MK to prioritize euploid embryos. This will be compared with a senior embryologist and a model built on live birth.

Declaration of interests: T.B. has nothing to disclose. R.S. has nothing to disclose. C.E. has nothing to disclose. R.D.S. has nothing to disclose. A.B. has nothing to disclose. S.M. has nothing to disclose. A.C. Is minor share holder of Care Fertility. A.C. has nothing to disclose.

REFERENCES

1. Gardner DK, Schoolcraft SW. In vitro culture of human blastocyst. In: Jansen R, Mortimer D, editors. Towards Reproductive Certainty: Infertility and Genetics Beyond 1999. The Plenary Proceedings of the 11th World Congress. Parthenon Publishing Group Inc; 1999:378–88.
2. Edwards RG, Purdy JM, Steptoe PC, Walters DE. The growth of human preimplantation embryos in vitro. Am J Obstet Gynecol 1981;141: 408–16.
3. Cummins JM, Breen TM, Harrison KL, Shaw JM, Wilson LM, Hennessey JF. A formula for scoring human embryo growth rates in in vitro fertilization: its value in predicting pregnancy and in comparison with visual estimates of embryo quality. J In Vitro Fert Embryo Transf 1986;3:284–95.
4. Bamford T, Barrie A, Montgomery S, Dhillon-Smith R, Campbell A, Easter C, Coomarasamy A. Morphological and morphokinetic associations with aneuploidy: a systematic review and meta-analysis. Hum Reprod Update 2022; 28:656–86.
5. Bamford T, Easter C, Montgomery S, Smith R, Dhillon-Smith RK, Barrie A, et al. A comparison of 12 machine learning models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos. Hum Reprod 2023;38:569–81.
6. Theobald R, SenGupta S, Harper J. The status of preimplantation genetic testing in the UK and USA. Hum Reprod 2020;35:986–98.
7. Cornelisse S, Zagers M, Kostova E, Fleischer K, van Wely M, Mastenbroek S. Preimplantation genetic testing for aneuploidies (abnormal number of chromosomes) in in vitro fertilisation. Cochrane Database Syst Rev 2020;9: CD005291.
8. Neal SA, Morin SJ, Franasiak JM, Goodman LR, Juneau CR, Forman EJ, et al. Preimplantation genetic testing for aneuploidy is cost-effective, shortens treatment time, and reduces the risk of failed embryo transfer and clinical miscarriage. Fertil Steril 2018;110:896–904.
9. Munné S, Kaplan B, Frattarelli JL, Child T, Nakhuda G, Shamma FN, et al. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. Fertil Steril 2019;112: 1071–9.e7.
10. Tiegs AW, Tao X, Zhan Y, Whitehead C, Kim J, Hanson B, et al. A multicenter, prospective, blinded, nonselection study evaluating the predictive value of an aneuploid diagnosis using a targeted next-generation sequencing–based preimplantation genetic testing for aneuploidy assay and impact of biopsy. Fertil Steril 2021;115:627–37.
11. Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo ranking intelligent classification algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reprod Biomed Online 2020;41:585–93.
12. Barnes J, Malmsten J, Zhan Q, Hajirasouliha I, Elemento O, Sierra J, et al. Noninvasive detection of blastocyst ploidy (euploid vs. aneuploid) using artificial intelligence (AI) with deep learning methods. Fertil Steril 2020; 114:e76.
13. Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can predict embryo ploidy status based on time-lapse data. Reprod Biol Endocrinol 2021;19:185.
14. Ortiz JA, Morales R, Lledó B, Vicente JA, González J, García-Hernández EM, et al. Application of machine learning to predict aneuploidy and mosaicism in embryos from in vitro fertilization cycles. AJOG Glob Rep 2022;2:100103.
15. Gazzo E, Peña F, Valdéz F, Chung A, Bonomini C, Ascenzo M, et al. The Kidscore™ D5 algorithm as an additional tool to morphological assessment and PGT-A in embryo selection: a time-lapse study. JBRA Assist Reprod 2020;24: 55–60.
16. Sawada Y, Sato T, Nagaya M, Saito C, Yoshihara H, Banno C, et al. Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. Reprod Biomed Online 2021;43:843–52.
17. Miyagi Y, Habara T, Hirata R, Hayashi N. Feasibility of deep learning for predicting live birth from a blastocyst image in patients classified by age. Reprod Med Biol 2019;18:190–203.
18. Huang TTF, Kosasa T, Walker B, Arnett C, Huang CTF, Yin C, et al. Deep learning neural network analysis of human blastocyst expansion from time-lapse image files. Reprod Biomed Online 2021;42:1075–85.
19. Huang B, Zheng S, Ma B, Yang Y, Zhang S, Jin L. Using deep learning to predict the outcome of live birth from more than 10,000 embryo data. BMC Pregnancy Childbirth 2022;22:36.
20. Pribenszky C, Nilselid AM, Montag M. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. Repro Biomed Online 2017; 35:511–20.
21. Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. Cochrane Database Syst Rev 2019;5:CD011320.
22. Kragh MF, Karstoft H. Embryo selection with artificial intelligence: how to evaluate and compare methods? J Assist Reprod Genet 2021;38:1675–89.
23. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370:1453–7.
24. Fertility 2021 barriers and breakthroughs 6-10th January 2021 Online. Hum Fertil (Camb) 2021;24:46–69.
25. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 2017;17: 162.
26. Peng C-YJ, Harwell M, Liou S-M, Ehman LH. Advances in missing data methods and implications for educational research. Real Data Anal 2006: 3178.
27. Horta F, Vollenhoven B, Healey M, Busija L, Catt S, Temple-Smith P. Male ageing is negatively associated with the chance of live birth in IVF/ICSI cycles for idiopathic infertility. Hum Reprod 2019;34:2523–32.
28. Vončina SM, Stenqvist A, Bungum M, Schyman T, Giwercman A. Sperm DNA fragmentation index and cumulative live birth rate in a cohort of 2,713 couples undergoing assisted reproduction treatment. Fertil Steril 2021;116:1483–90.
29. Kushnir VA, Frattarelli JL. Aneuploidy in abortuses following IVF and ICSI. J Assist Reprod Genet 2009;26:93–7.
30. Soler A, Morales C, Mademont-Soler I, Margarit E, Borrell A, Borobio V, et al. Overview of chromosome abnormalities in first trimester miscarriages: a series of 1,011 consecutive chorionic villi sample karyotypes. Cytogenet Genome Res 2017;152:81–9.
31. Chawla M, Fakih M, Shunnar A, Bayram A, Hellani A, Perumal V, et al. Morphokinetic analysis of cleavage stage embryos and its relationship to aneuploidy in a retrospective time-lapse imaging study. J Assist Reprod Genet 2015;32:69–75.

32. Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Hickman CF. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. Reprod Biomed Online 2013;26:477–85.
33. Basile N, Nogales C, Bronet F, Florensa M, Riqueiros M, Rodrigo L, et al. Increasing the probability of selecting chromosomally normal embryos by time-lapse morphokinetics analysis. Fertil Steril 2014;101:699–704.
34. Mumusoglu S, Ozbek IY, Sokmensuer LK, Polat M, Bozdag G, Papanikolaou E, Yarali H. Duration of blastulation may be associated with ongoing pregnancy rate in single euploid blastocyst transfer cycles. Reprod Biomed Online 2017;35:633–9.
35. Del Carmen Nogales M, Bronet F, Basile N, Martinez EM, Liñán A, Rodrigo L, Meseguer M. Type of chromosome abnormality affects embryo morphology dynamics. Fertil Steril 2017;107:229–35.e2.
36. Desai N, Ploskonka S, Goodman LR, Austin C, Goldberg J, Falcone T. Analysis of embryo morphokinetics, multinucleation and cleavage anomalies using continuous time-lapse monitoring in blastocyst transfer cycles. Reprod Biol Endocrinol 2014;12:54.
37. Kovacs P, Matyas S, Forgacs V, Sajgo A, Rarosi F, Pribenszky C. Time-lapse embryo selection for single blastocyst transfer – results of a multicenter, prospective, randomized clinical trial. Fertil Steril 2013;100:S90.
38. Kahraman S, Çetinkaya M, Pirkevi C, Yelke H, Kumtepe Y. Comparison of blastocyst development and cycle outcome in patients with eSET using either conventional or time lapse incubators. A prospective study of good prognosis patients. J Reprod Stem Cell Biotechnol 2012;3:55–61.
39. Siristatidis C, Komitopoulou MA, Makris A, Sialakouma A, Botzaki M, Mastorakos G, et al. Morphokinetic parameters of early embryo development via time lapse monitoring and their effect on embryo selection and ICSI outcomes: a prospective cohort study. J Assist Reprod Genet 2015;32:563–70.
40. Meng Q, Xu Y, Zheng A, Li H, Ding J, Xu Y, et al. Noninvasive embryo evaluation and selection by time-lapse monitoring vs. conventional morphologic assessment in women undergoing in vitro fertilization/intracytoplasmic sperm injection: a single-center randomized controlled study. Fertil Steril 2022;117:1203–12.
41. Goodman LR, Goldberg J, Falcone T, Austin C, Desai N. Does the addition of time-lapse morphokinetics in the selection of embryos for transfer improve pregnancy rates? A randomized controlled trial. Fertil Steril 2016;105:275–85.e10.
42. Rubio I, Galán A, Larreategui Z, Ayerdi F, Bellver J, Herrero J, et al. Clinical validation of embryo culture and selection by morphokinetic analysis: a randomized, controlled trial of the embryoscope. Fertil Steril 2014;102:1287–94.e5.
43. Ahlström A, Lundin K, Lind AK, Gunnarsson K, Westlander G, Park H, et al. A double-blind randomized controlled trial investigating a time-lapse algorithm for selecting Day 5 blastocysts for transfer. Hum Reprod 2022;37:708–17.
44. Ciray HN, Aksoy T, Goktas C, Ozturk B, Bahceci M. Time-lapse evaluation of human embryo development in single versus sequential culture media–a sibling oocyte study. J Assist Reprod Genet 2012;29:891–900.
45. Munoz Munoz E, Ojeda M, Perez M, Taboas E, Meseguer M, Aguilar J, editors. Morphology dynamics of multinucleation affects embryo chromosome content. Hum Reprod. Abstract book of the 30 th ESHRE Annual Meeting, Munich, Germany, 29 June – 2 July 2014;29:i1–389.
46. Lemmen JG, Agerholm I, Ziebe S. Kinetic markers of human embryo quality using time-lapse recordings of IVF/ICSI-fertilized oocytes. Reprod Biomed Online 2008;17:385–91.
47. Agerholm IE, Hnida C, Crüger DG, Berg C, Bruun-Petersen G, Kølvraa S, Ziebe S. Nuclei size in relation to nuclear status and aneuploidy rate for 13 chromosomes in donated four cells embryos. J Assist Reprod Genet 2008;25:95–102.
48. Bellver J, Mifsud A, Grau N, Privitera L, Meseguer M. Similar morphokinetic patterns in embryos derived from obese and normoweight infertile women: a time-lapse study. Hum Reprod 2013;28:794–800.
49. Fréour T, Dessolle L, Lammers J, Lattes S, Barrière P. Comparison of embryo morphokinetics after in vitro fertilization-intracytoplasmic sperm injection in smoking and nonsmoking women. Fertil Steril 2013;99:1944–50.
50. Kirkegaard K, Hindkjaer JJ, Ingerslev HJ. Effect of oxygen concentration on human embryo development evaluated by time-lapse monitoring. Fertil Steril 2013;99:738–44.e4.

CHAPTER 5: A COMPARISON OF MORPHOKINETIC MODELS AND MORPHOLOGICAL SELECTION FOR PRIORITISING EUPLOID EMBRYOS: A MULTICENTRE COHORT STUDY

Thomas Bamford, Rachel Smith, Selina Young, Amy Evans, Megan Lockwood, Christina Easter, Sue Montgomery, Amy Barrie, Rima Dhillon-Smith, Arri Coomarasamy, Alison Campbell

Manuscript accepted for publication in Human Reproduction on 2nd November 2023

My role in this publication is as follows: I collected all cleaned all the data used in this study. I organised and arranged a team of embryologist to assist with the ranking of embryos. I performed all statistical analyses and wrote the whole of the manuscript.

Abstract

Study Question: Are morphokinetic models better at prioritising a euploid embryo for transfer over morphological selection by an embryologist?

Summary Answer: Morphokinetic algorithms lead to an improved prioritisation of euploid embryos when compared to embryologist selection.

What is known already: PREFER (Predicting euploidy for embryos in reproductive medicine) is a previously published morphokinetic model associated with live birth and miscarriage. The second model uses live birth as the target outcome (LB model).

Study design: Data for this cohort study were obtained from 1958 biopsied blastocysts at nine IVF clinics across the UK from January 2021 to December 2022.

Participants/Materials, setting, method: The ability of the PREFER and LB models to prioritise a euploid embryo was compared against arbitrary selection and the prediction of four embryologists using the timelapse video, blinded to the morphokinetic time stamp. The comparisons were made using calculated percentages and normalised discounted cumulative gain (NDCG), whereby an NDCG score of 1 would equate to all euploid embryos being ranked first. In arbitrary selection, the ploidy status was randomly assigned within each cycle and the NDGC calculated, and this was then repeated 100 times and the mean obtained.

Main results and role of chance: Arbitrary embryo selection would rank a euploid embryo first 37% of the time, embryologist selection 39%, and the LB and PREFER ploidy morphokinetic models 46% and 47% of the time, respectively. The AUC for LB and PREFER model was 0.62 and 0.63, respectively. Morphological selection did not significantly improve the performance of both morphokinetic models when used in combination. There was a significant difference between the NDGC metric of the PREFER model versus embryologist selection at 0.96 and 0.87, respectively ($t=14.1$, $p<0.001$). Similarly, there was a significant difference between the LB model and embryologist selection with a NDGC metric of 0.95 and 0.87, respectively ($t=12.0$, $p<0.001$). All four embryologists ranked embryos similarly, with an intraclass coefficient of 0.91 (95% CI 0.82-0.95, $p<0.001$).

Limitations and reason for caution: Aside from the retrospective study design, limitations include allowing the embryologist to watch the time lapse video, potentially providing more information than a truly static morphological assessment. Furthermore, the embryologists at the participating centres were familiar with the significant variables in time lapse, which could bias the results.

Wider implications of findings: The present study shows that the use of morphokinetic models, namely PREFER and LB, translates into improved euploid embryo selection.

Study funding and competing interests: This study received no specific grant funding from any funding agency in the public, commercial or not-for-profit sectors. Dr Alison Campbell is minor share holder of Care Fertility. All other authors have no conflicts of interest to declare. Time lapse is a technology for which patients are charged extra at participating centres.

Introduction

Embryo selection is paramount to the success of assisted conception. Morphological assessment is known to be associated with clinical outcome, and therefore is commonly used in IVF laboratories worldwide for embryo selection, utilising assessment criteria, such as Gardner's (Gardner et al., 2000). Historically, it has been reported that there may be a high degree of inter-observer and intra-observer variability using this method (Arce et al., 2006, Baxter Bendus et al., 2006, Paternot et al., 2011). Therefore, the concept of using an enclosed incubation time lapse device, allowing the precise annotation of an embryo's development seemed like a breakthrough for the field. Particularly since this offered a great deal more information over static assessment, allowing predictive models to be created based upon morphokinetic variables. Meseguer et al. (2011) proposed the first hierarchical morphokinetic model for predicting live birth; since this model, several others have been proposed (Fréour et al., 2013, Huang et al., 2022, Meseguer et al., 2011). Many similar models are now commercially available, yet there is significant disagreement amongst IVF professionals as to whether time lapse incubation confers any improvement in live birth rates per embryo transfer. Furthermore, there is also significant heterogeneity in the efficacy of different morphokinetic models, possibly accounting for the contradictory reports of clinical effectiveness (Bamford et al., 2023, Barrie et al., 2017, Storr et al., 2018).

In order to try and improve the accuracy, some studies have used ploidy status as the target outcome despite the controversies surrounding pre-implantation genetic testing for aneuploidy (PGT-A) (Basile et al., 2014, Campbell et al., 2013, Chawla et al., 2015, Del

Carmen Nogales et al., 2017, Desai N., 2016). Arguably, this may be beneficial since ploidy status is less affected by confounding variables when compared to live birth, for example, BMI, parity, laboratory factors, and endometrial thickness. Furthermore, aneuploid embryos have a global delayed development in comparison to their euploid counterparts, possibly due to changes in complex biochemical processes that occur when errors have been detected by the developing embryo (Bamford et al., 2022, Coticchio et al., 2021). This therefore allows the embryos to be stratified for risk of aneuploidy based upon their morphokinetic timings using complex statistical modelling. This may be beneficial since there are now over 100 studies which suggest improved clinical outcomes following PGT-A (Griffin, 2022). For instance, one study examined 2464 PGT-A cycles taken from Human Fertilisation and Embryology Authority data (2016-2018); all age groups showed a significantly higher live birth rate per embryo transfer or shortened time to pregnancy with the use of PGT-A (Sanders et al., 2021). Despite this, there are several randomised controlled trials (RCTs) suggesting that PGT-A is ineffective at improving clinical outcomes; this is largely because PGT-A is ineffective at improving cumulative pregnancy rates (Munné et al., 2019, Verpoest et al., 2018, Yan et al., 2021). For instance, when subgroups of older women are analysed within these trials, the live birth rates per embryo transfer are significantly higher with PGT-A; therefore, there may be limitations in the design and the selection criteria of published RCTs. Controversies aside, it is indisputable that the use of PGT-A is increasing worldwide (Theobald et al., 2020). Therefore, finding less invasive alternatives has become an important aim for many researchers.

The use of artificial intelligence (AI) within the embryology laboratory has gained enormous traction in recent years. It has potential advantages over traditional statistical modelling as it allows the exploitation of interconnections between predictors while also learning from incorrect classifications. This technology is therefore well suited to embryo selection, particularly in cases of blastocyst image analysis (Chavez-Badiola et al., 2020, Tran et al., 2019). An investigation was therefore justified to determine whether AI was superior to traditional statistics in the case of numerical analysis of morphokinetic variables (Bamford et al., 2023). Twelve morphokinetic models were built by our group, using four different algorithms on a dataset of 8147 biopsied blastocysts. It was concluded that a model based on logistic regression was the best performing algorithm for predicting ploidy status. This was not surprising since AI typically requires hundreds of thousands of data points, such as that provided by image analysis. Additionally, we concluded that clinical variables, such as age, should not be included into morphokinetic models. While age may be a strong predictor of ploidy status, this is not useful at the level of the embryo. The strength of this association means that the model places a larger amount of weight to this predictor, such that when embryos from patients of same age or from the same patient are ranked, there is little variability in the model risks scores. This results in the model ineffectively ranking embryos as all, or most, are given a similar score. A ‘morphokinetics only’ approach was therefore more effective as it takes its predictions solely from embryo quality; this was named PREFER (Predicting euploid for embryos in reproductive medicine).

The association between PREFER and miscarriage and live birth was then investigated using a separate dataset of 3587 single embryo transfers (Bamford et al. 2023b). The PREFER risk score was significantly associated with live birth and miscarriage; for example, an embryo deemed ‘low risk’ by PREFER was significantly more likely to result in a live birth than those ‘high risk’ (OR 1.95; 95% CI 1.65-2.25; $p<0.001$). While these results demonstrate an association, it is not yet clear whether this model is superior at prioritising a euploid embryo compared to morphological ranking by an embryologist. Those against the use of time lapse culture would argue that there are now several RCTs that have demonstrated it results in no improvement in live birth rate (Ahlström et al., 2022, Armstrong et al., 2019). Our study therefore aimed to determine whether a morphokinetic model prioritised embryos for transfer that were more likely to be euploid when compared to embryologists using morphological assessment of timelapse videos. Two models were compared; PREFER, and a morphokinetic model used in the participating centres to predict live birth (LB model).

Materials and Methods

Data for this cohort study were obtained from 1958 biopsied blastocysts at nine IVF clinics across the UK from 2021 to 2022. The participating centres are part of a fertility group providing private and state funded treatments. Each centre conforms to the same laboratory practices. There were no changes in laboratory procedures during the study period. This study was granted institutional review board approval (CARE/ERC/09.02.2023). This research did not receive any grant from funding agencies

in the public, commercial, or not-for-profit sectors. Participants consent to their data being used for research purposes as part of their treatment at participating centres.

Participants

Women were included in this study if they had an autologous PGT-A cycle with the use of time lapse. Embryos that contributed data to previous model derivation were not used this study. Cycles with only one biopsied embryo were excluded as this study aimed to determine the ability of morphokinetic models to rank and prioritise euploid embryos. A total of 1608 biopsied blastocysts remained from 498 patients who had PGT-A for advanced maternal age, recurrent implantation failure (≥ 3 failed embryo transfers), recurrent miscarriage (≥ 3 miscarriages) or to shorten the time to pregnancy. Patients underwent pituitary suppression and ovarian stimulation using either the long GnRH agonist or short antagonist protocol. Transvaginal oocyte retrieval was performed under sedation 36 hours after the trigger injection (hCG or agonist) when three follicles reached 17mm or more. All embryos were cultured in a time lapse system, EmbryoScope (Vitrolife, Frölunda Sweden) at 37°C, 6-6.5% CO_2, 5% O_2 and 88.5-89% N_2. The wells of the Embryoscope slide (Vitrolife, Frölunda, Sweden) were filled with culture medium Global total LP (Cooper Surgical; Trumbull, CT, USA) and overlaid with LifeGuard mineral oil (Cooper Surgical; Trumbull, CT, USA). Morphokinetic parameters were manually annotated by an embryologist trained in house, following published guidelines and utilising an in-house quality assurance process (Barrie et al., 2021, Ciray et al., 2014). Every 10 minutes microscopic images were acquired of the embryos through seven multifocal planes. All blastocysts underwent laser assisted hatching on

day 3 to facilitate trophectoderm herniation, and it has previously been reported that this has no effect on time from insemination to start of blastulation (tSB) or formation of a full blastocyst (tB) (Campbell et al., 2013). Those embryos suitable for biopsy on day 5 or 6, had 5-10 cells aspirated as described elsewhere (Campbell et al., 2013). The samples were analysed using next-generation sequencing (NGS, Cooper Surgical, Trumbull, CT, USA). Embryo grading was performed according to Gardner's classification (Gardner D. K., 1999).

Model development

Two morphokinetic models were compared, with one using ploidy status as the target outcome (PREFER) and the other using live birth (LB model). Morphokinetic model development has been extensively described elsewhere (Bamford, et al., 2023). For the model, a sample of 8147 biopsied blastocysts was used to create a mixed effects logistic regression model using the patient identifier as the random intercept, allowing for within-cluster variability. Predictors were selected for inclusion in the model using backward elimination with a p value <0.157. The final PREFER model included: time to 6 cells (t6), 7 cells (t7), start of compaction (tSC), formation of morula (tM), tB. This model had a meta-analysed AUC of 0.61 and F1 score of 0.72, calculated using internal-external cross validation (Bamford et al., 2023). The LB model had AUC of 0.68 and was built using the same methodology with a sample of 6228 fresh embryo transfers. This LB model is an updated version of that reported by Fishel et al (2018), and the latest algorithm has been validated in-house only. The following final variables were included in the LB model: tSB, trophectoderm and morula grade, the interval between tB and tSB, and kinetic interval

calculations of t3, t4, t5, t8. The PREFER model categorised embryos into six risk scores with a score of 1 indicating a high risk of aneuploidy and a score of 6, more likely euploid. The LB model was constructed such that an embryo was categorised into one of 10 scores, with score 1 indicating the lowest, and score 10 associated with the highest chance of live birth.

Sample size

The sample size for this study was chosen to ensure that estimates of accuracy were made with adequate precision. We deemed a sensitivity and specificity of 70% to be clinically useful. We chose a sample size of 1200 embryos, using a prevalence of euploid embryos of 37%. The CIs of 70% for sensitivity would be 65.5%-72.3% and 66.6-73.2% for specificity.

Statistical analysis

The association between risk scores for both models and ploidy status was investigated through graphical analysis and percentages. This was performed for all patients and a subgroup analysis of different patient age groups including, ≤35, 36-38 and >38 years. Four embryologists were asked to rank a random sample of 100 embryos within each treatment cycle from best prognosis (1) to worst (‘n’), where ‘n’ is the number of blastocysts in the cycle. The embryologist was permitted to watch the time lapse video as often as they wanted and the morphokinetic timestamp was edited out of the video so that the embryologist was blinded to morphokinetic timings. Intraclass correlation coefficient (ICC) is an index used to reflect the degree of correlation and agreement

between measurements, whereby values of <0.5 indicate poor reliability and values >0.9 indicate excellent reliability. All four embryologists ranked embryos similarly, with an ICC of 0.91 (95% CI 0.82-0.95, p<0.001). Each embryologist had at least 5 years of experience. This ranking was compared to that provided by the LB and PREFER models. Where the morphokinetic models had more than one embryo with the same score, they were denoted as the same rank.

A statistical metric called normalised discontinued cumulative gain (NDCG) was used to determine the accuracy of the rank for ploidy prediction. It uses a weighted scale based upon the relative position in the list (Jarvelin and Kekalainen, 2002). Should all euploid embryos be placed at the top of the rank, it would score a 1. The mean NDGC score was compared for each embryo selection technique investigated in the sample of 100 embryos that the embryologists were asked to rank. The Student's t-test was utilised to determine whether there was statistically significant difference between the approaches. Each selection technique was also compared to arbitrary embryo selection, where ploidy status was randomly assigned within each cycle and the NDGC calculated, and this was then repeated 100 times and the mean obtained. This was undertaken using a random number processor or in Python. A final approach was compared for each model, whereby for embryos with equal ranks, morphological grading was used to prioritise the better quality embryo. All variables required to compute the risk scores had <5% missing data and therefore they were considered missing at random; no imputation was required.

Model performance was also analysed using the metrics: AUC, accuracy, positive predictive value, negative predictive value, false positive rate, false negative rate, sensitivity, specificity, precision and F1 score. It must be considered that this required the model to use a defined threshold of predicted probability to classify an embryo as ‘euploid’ or ‘aneuploid’. The models would never be expected to determine this, only the risk of aneuploidy; however, such metrics provide important information about performance.

Results

An overview of patient characteristics is provided in Table 1. There was an association between a lower PREFER model risk score i.e., worse prognosis, and a higher proportion of aneuploid embryos (Fig. 1). In contrast, the trend in the proportion of euploid and aneuploid embryos for the LB model risk scores is less clear. The general trend between a higher PREFER score and proportion of euploid embryos remained for each subgroup of ages (Supplementary Fig. S1). There was a clearer relation demonstrated for the LB model risks score and percentage of euploid embryos in 36- to 38-year-olds compared to all other age groups and all ages combined (Supplementary Fig. S1). Furthermore, there is a correlation between the morphological grade of the embryo (as assessed by an embryologist) and the PREFER and LB model probabilities of euploidy and live birth, respectively (Fig. 2).

Figure 1: Percentage of euploid and aneuploid embryos in each risk score of the LB model and PREFER model

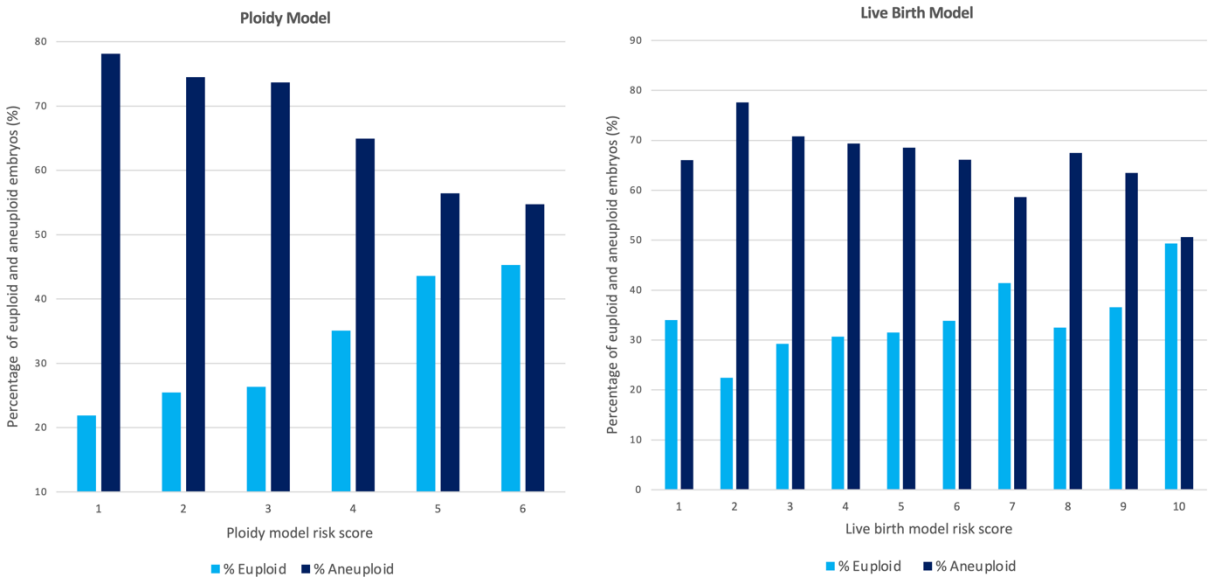


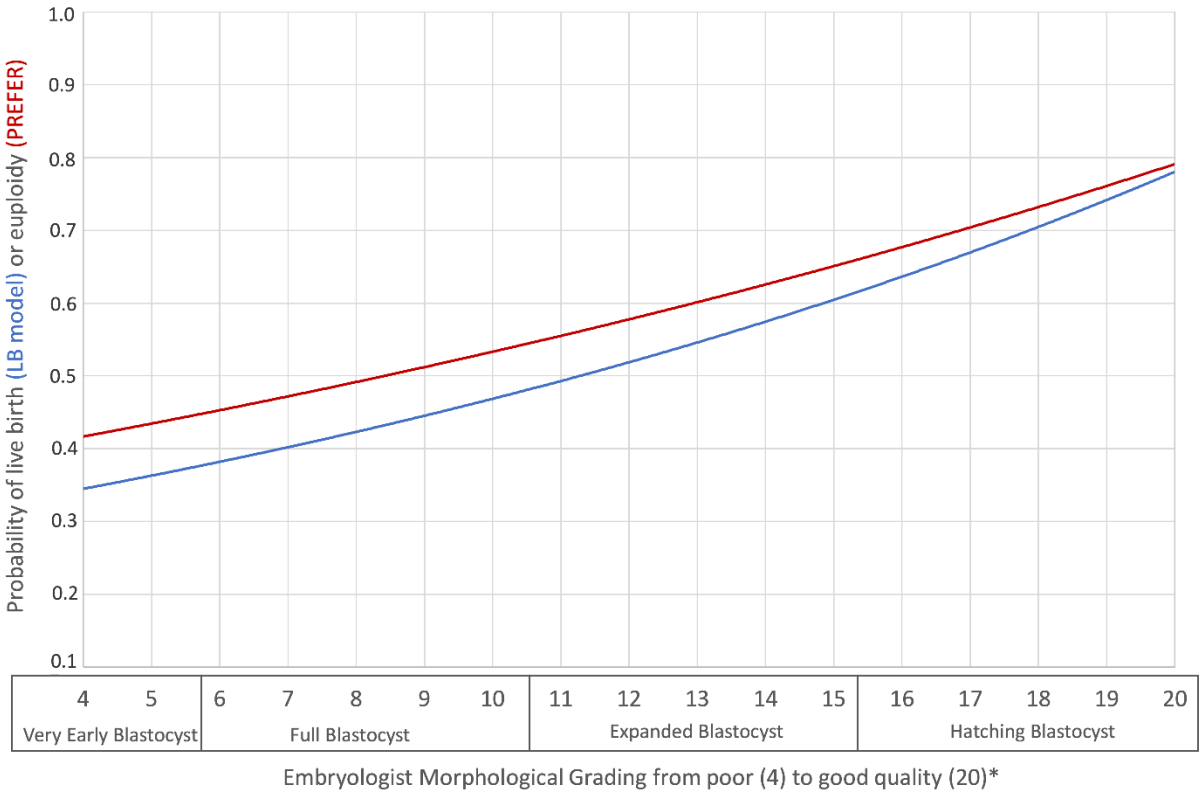
Table 1: Participant Characteristics (n=1608 embryos)

	Euploid	Aneuploid	p-value
Number of embryos (%)	617 (41%)	991 (59%)	
Mean oocyte provider age, (SD)*	36.7 (3.44)	38.6 (3.58)	<0.001
Oocyte provider age**			
<35	836 (52%)	772 (48%)	<0.001
35-37	800 (50%)	808 (50%)	
38-39	563 (35%)	1045 (65%)	
40-42	434 (27%)	1174 (73%)	
>=43	61 (9.5%)	1463 (91%)	
BMI, Median (IQR)*	23.9 (21.4-25.9)	23.3(21.1-27.1)	0.785
AMH, Median (IQR)*	15.9 (10.1-29.1)	15.1 (9.1-26.2)	0.004
AFC, Median (IQR)*	10 (0-15)	10 (0-16)	0.825
Number of oocytes retrieved, Median (IQR)*	14 (10-18)	13 (8-15)	<0.001
Sperm provider age, Mean (SD)*	38.5 (5.11)	40.6 (5.83)	<0.001
Protocol type**			
			<0.001
Short	1383 (86%)	1303 (81%)	
Long	225 (14%)	305 (19%)	
Average FSH dose received**			
75-150 IU	772 (48%)	836 (52%)	
187.5-225 IU	595 (37%)	1013 (63%)	<0.001
262.5-337.5 IU	547 (34%)	1061 (66%)	
>=450 IU	434 (27%)	1174 (73%)	
Embryos created from donated sperm*	50 (3.1%)	87 (5.4%)	0.001
Grading at biopsy**			
Full Blastocyst	402 (25%)	450 (28%)	<0.001
Expanded Blastocyst	225 (14%)	338 (21%)	
Hatching Blastocyst	981 (61%)	820 (51%)	
ICM Grade**			
1	338 (21%)	241 (15%)	0.135
2	1158 (72%)	1158 (72%)	
3	113 (7%)	209 (13%)	
Trophectoderm Grade**			
1	193 (12%)	209 (13%)	0.025
2	997 (62%)	932 (58%)	
3	418 (26%)	466 (29%)	
Morula Grade**			
1	756 (47%)	740 (46%)	0.432
2	852 (53%)	868 (54%)	
ART method (n, %)**			
IVF	868 (54%)	820 (51%)	0.826
ICSI	740 (46%)	788 (49%)	
Sperm Concentration (median, IQR)*	65 (36-93)	63 (34-91)	0.401

*Man-Whitney U Test for statistical significance
** Chi^2 test for significance

Key- SD- Standard deviation, BMI- body mass index, AMH- Anti-Mullerian hormone, AFC- antral follicle count, IQR- interquartile range, FSH- follicle stimulating hormone, ICM- inner cell mass, ART- assisted reproductive technology.

Figure 2: Graph to show the correlation between embryologist morphological ranking and probability of euploid and live birth as predicted by PREFER and LB model, respectively (n=783 embryos). Key in footnote below.



**4=1AB/1BC/2AB/2BC, 5=1AA/2AA, 6=3CC, 7=3BC/3CB, 8=3BB/3AC/3CA, 9=3AB/3BA, 10=3AA, 11=4CC, 12=4BC/4CB, 13=4BB/4AC/4CA, 14=4AB/4BA, 15=4AA, 16=5CC/6CC, 17=5BC/5CB/6BC/6CB, 18=5BB/5AC/5CA/6BB/6AC/6CA, 19=5AB/6AB/5BA/6BA, 20=5AA/6AA as per Gardner's criteria*

The abilities of the PREFER and LB models and four embryologists to prioritise euploid embryos were compared using the NDGC metric (Table 2). Arbitrary embryo selection would rank a euploid embryo first 37% of the time. The embryologist, LB model and PREFER model ranked a euploid embryo first 39%, 46% and 47% of the time, respectively. Morphological selection did not significantly improve the performance of either morphokinetic model when used in combination. There was a significant difference between the NDGC metric of the PREFER model versus embryologist selection, at 0.96 and 0.87, respectively ($t=14.1$, $p<0.001$). Similarly, there was a significant difference between the LB model and embryologist selection with a NDGC metric of 0.95 and 0.87, respectively ($t=12.0$, $p<0.001$). All four embryologists ranked embryos similarly, with an ICC of 0.91 (95% CI 0.82-0.95, $p<0.001$). Table 3 outlines model performance metrics for the LB and PREFER models: model performance was similar (AUC=0.62 and 0.63, respectively). Both models are sensitive but not specific at detecting an aneuploid embryo.

Table 2: Results of NDCG analysis, comparing ranking of four embryologists, the live birth and PREFER model and finally the respective models combined with morphological grading

	Random	Embryologist	Live Birth Model	PREFER Model	LB Model + morphological grading	PREFER model + morphological grading
Mean NDCG* ***Standard Deviation***	0.85 *(SD 0.01)*	0.87 *(SD 0.047)*	0.95 *(SD 0.085)*	0.96 *(SD 0.079)*	0.96 *(SD 0.069)*	0.96 *(SD 0.059)*
Euploid embryo in the top of the rank	37%	18/46 39%	173/376 46%	188/399 47%	134/280 48%	136/280 49%

** A statistical metric called normalised discontinued cumulative gain (NDCG) was used to determine the accuracy of the rank for ploidy predication. This uses a weighted scale based upon the relative position in the list (Jarvelin and Kekalainen, 2002). Should all euploid embryos be placed at the top of the rank, it would score a 1. The mean NDGC score was compared for each embryo selection technique investigated in the sample of 100 embryos that the embryologists were asked to rank.*

Table 3: Model Performance Metrics

	LB model	PREFER
AUC	0.62	0.63
Accuracy (%)	63	60
Positive predictive value (%)	58	62
Negative predictive value (%)	45	48
False positive rate (%)	60	62
False negative rate (%)	18	15
Sensitivity (%)	72	75
Specificity (%)	28	31
Precision (%)	58	55
F-1 score*	0.68	0.69

**F-1 score is a statistic used in machine learning to measure the balance between precision and recall*

Discussion

This multicentre cohort study demonstrated that the use of morphokinetics resulted in an improved prioritisation of euploid embryos when compared to selection of blastocysts by an embryologist, without the use of a morphokinetic model. Two morphokinetic models were compared, one designed to select embryos based upon ploidy status (PREFER) and another built with live birth as the target outcome (LB model). PREFER and the LB model improved the chance of ranking a euploid embryo at the top when compared to arbitrary selection (46% and 47%, respectively versus 37%). Whilst a variety of performance metrics have also been reported in Table 3, this study focuses on the ranking ability of these algorithms. This is because asking a morphokinetic model to classify an embryo as ‘euploid’ or ‘aneuploid’ is not representative of how these models will be expected to function when selecting embryos. The aim is for morphokinetic models to determine chance of euploidy or live birth, but they are not intended to be a diagnostic tool.

It was investigated whether morphokinetic models aid decision making by using morphology grading to decide between tied ranks, however this did not demonstrate any improvement in the prioritisation of a euploid embryo. This may be because there is significant collinearity between morphology and morphokinetics; such that morphology adds little prognostic information in the context of morphokinetics which is more quantitative and potentially discriminatory. This has been demonstrated in Fig. 2, whereby a higher grade embryo is also correlated with a better prediction of live birth and ploidy

status by a morphokinetic model. This is not unsurprising since we know that morphological assessment of embryos, whist subjective, is associated with live birth, and therefore by association, euploidy. A previous study and Fig. 1 illustrate the association between ploidy status and PREFER (Bamford et al., 2023). The present study therefore concludes that the use of morphokinetic models, namely PREFER and LB, translates into improved euploid embryo selection.

The ability of morphokinetic algorithms to discriminate between euploid and aneuploid embryos remains a controversial area despite a decade of research (Campbell, Fishel, Bowman, Duffy, Sedler and Hickman, 2013, Chawla et al., 2015, Del Carmen Nogales et al., 2017, Desai et al., 2014, Huang et al., 2021, Mumusoglu et al., 2017, Uyar et al., 2015). More recent research has failed to clarify this issue; Quinn et al. (2022) reported that morphokinetics did not distinguish between euploid and aneuploid embryos. However, little can be extrapolated from the Quinn study since only low-quality embryos (n=328) were examined and no algorithm was created; only individual parameters were examined. This is in contrast to the findings from a large systematic review and meta-analysis on this topic (Bamford et al., 2022). In contrast, a well-designed study by Kato et al. (2023) included 3573 biopsied blastocysts and reported that euploidy was significantly correlated with two commercially available live birth prediction morphokinetic models, iDAS and KIDScore. Their reported findings are concordant with those reported in this manuscript, and we also agree that while there is an association with ploidy status, no morphokinetic model should as yet be used as a diagnostic tool but it may be used for the prioritisation for biopsy or, for instance, where legislation prohibits embryo biopsy. This

may change with improvements to ploidy prediction models, perhaps with the addition of blastocyst image, video analysis, proteomics and metabolics (Chavez-Badiola et al., 2020, Katz-Jaffe and McReynolds, 2013, Krisher et al., 2015, Payá Bosch et al., 2023, Tran et al., 2019).

When investigating time lapse, it must be considered that the benefits may arise from undisturbed culture or enclosed incubation rather than, or in addition to, the morphokinetic models themselves. Considering this, there have only been three RCTs which have investigated the use of time lapse compared to morphology while incubating all embryos, including those in the control group, in a time lapse device (Ahlström et al., 2022, Goodman et al., 2016, Kaser et al., 2017). While these studies do not investigate ploidy prediction models, they do provide some indication on the clinical utility of morphokinetics as one can extrapolate that a model built on live birth also targets euploidy since most live births will be euploid. All of these studies concluded that time lapse did not lead to significant improvement in clinical outcomes. Although, it must be noted that in the study by Goodman et al. (2016), only 75% of randomised patients underwent blastocyst transfer. Furthermore, Kaser et al. (2017) terminated the study after only 60% of patients had been randomised and Ahlström et al. (2022) randomised only half (776/1656), however, the latter study was considerably larger than prior RCTs. This meant that Ahlström et al. (2022) were only able to detect a difference of 10% at a power of 80%, compared to a 90% power for PREFER (Bamford et al. 2023b). In the study by Ahlström et al. (2022), the ongoing pregnancy rate was comparable in the two groups (47.4% in the time lapse group and 48.1% in the control). There was also no difference in ongoing

pregnancy after the adjustment of confounders. These three studies also investigated three different models, Kaser et al. (2017) used Early embryo viability assessment™, Ahlström *et al.* (2022) used KIDSCORE™ D5 and Goodman et al. (2016) investigated a selection algorithm developed in house. One could argue that it is therefore difficult to compare the results of one study to another as the interventions in these groups are not consistent, and the models tested were commercially available, generic models. Rubio et al. (2014) performed a large group-wide RCT (n=843) demonstrating a significant increase in ongoing pregnancy rate in the time lapse group compared to the control (51.4% versus 41.7%, p=0.005), however, the results of this study are difficult to interpert since the embryos in the control group were cultured in a standard.

While there have been several retrospective studies demonstrating positive associations between morphokinetic models and live birth, many studies have suggested that this does not translate into an improvement in clinical outcomes when compared to morphological selection (Kato et al., 2021, Lee et al., 2019, Reignier et al., 2018). Given that many morphokinetic models performed in a similar fashion to morphological selection in most trials, there remains an argument that while it does not conclusively confer an improvement in clinical outcomes, it may improve labaratory work flows, standardise selection and provide extra information over static assessment. It is this more detailed assessment that some professionals argue leads to improved selection in a smaller group of embryos displaying abnormal cleavages or other unusual phenomena, and is the basis of the reported success of such in-house developed and tested algorithms, as described in this study.

There are significant limitations to the design of this study that should be considered when interpreting the findings. Firstly, we allowed the embryologist to watch the whole time lapse video, and whilst this did not include the morphokinetic time stamp or morphokinetic annotations, it does provide more information than a truly static morphological assessment. Second, the embryologists at the participating centres were familiar with working with time laspe and had previously participated in quality assurance in time lapse annotation and embryo assessment. They had a good knowledge of the most predictive morphokinetic components used in the LB model, primarily, time to start blastulation (tSB), which has recently been reported to be the single most powerful morphokinetic variable for prediction of live birth (Campbell et al., 2022). Therefore, having the ability to repeatedly view the video of the developing embryo could introduce bias since the embryologists may be able to deterine embryos with a shorter tSB, albiet less accurately than with a time lapse device. That said, during the study design we wanted to be as fair as possible, hypothesising that morphokinetic models have an additive effect over and above the consideration of sole parametres, such as tSB. The alternative is to use a static image of a blastocyst, a metholodology used in the validation study by Chavez-Badiola et al. (2020). The limitation of this approach is that it doesn't provide a comparable amount of information for an embryologist to make a selection compared to the time-lapse video used in a real life.

It is known that many factors contribute to embryo quality over and above the ploidy status of the embryo (Coticchio et al, 2021). Therefore, it may be that a model built on live birth, while not more likely to select a euploid embryo over PREFER, may select an embryo

with improved viability. It could be argued that using the outcome of live birth is superior to ploidy status as it is a confirmed outcome, whereas ploidy status is inferred from a biopsy result. We must, therefore, be confident of the concordance of PGT-A biopsy results and true ploidy status of the embryo, particularly considering mosaic results. That said, studies have demonstrated reliable reporting, with no live births from aneuploid transfers in non-selection studies (Tiegs et al., 2021). This can only be truly tested by an RCT; considering the limitations of previous studies it would be important to achieve the desired randomisation number in order to be confident in the conclusions. Finally, we do not know whether the use of deep learning image analysis and automated annotation would improve the accuracy of morphokinetic models, and this will be explored in future studies.

One limitation of existing morphokinetic models, even those commercially available, is the lack of well-designed validation studies using robust methodology. Therefore, a strength of the present study is that the morphokinetic models have been externally and extensively validated and published in peer reviewed journals (Bamford et al., 2023a, Bamford et al., 2023b, Fishel et al., 2018). We therefore can be confident that the models have been designed to detect meaningful differences between embryos utilising very large datasets, and that results from these models are entirely consistent and reproducible. Second, the authors of this study have previously validated the annotation practices of embryologists in participating centres, finding that there was no significant difference between practitioners, as replicated in this study's findings (Barrie, 2021). A third strength is the size of the sample size to validate these models, however the sample

size was more limited in the group of embryos where embryologist ranked a cohort of each patient's embryos (n=46 patients, 100 embryos). Therefore, despite a negative trend, it might be that this study is underpowered to detect a difference between embryologist and morphokinetic model selection.

Conclusion

In conclusion, in this study it has been demonstrated that morphokinetic models offer an advantage in prioritising euploid embryos for transfer over selection by an embryologist alone. Two different approaches to morphokinetic models were compared, one aiming to predict ploidy status (PREFER) and another live birth (LB model). Whether the LB model, PREFER or morphological selection is better at improving clinical outcomes needs to be investigated further. This study forms the foundation for a future prospective RCT to include a specific investigation of a ploidy prediction algorithm, as that does not currently exist in the literature.

CHAPTER 6: THESIS CONCLUSIONS

The study of morphokinetics using a time-lapse device has given fertility practitioners a wealth of information about embryo development. Indeed, phenomena have been observed using these enclosed incubation systems that would be more challenging to study in standard incubation. What is less clear, is whether the lessons learnt and morphokinetic algorithms created using annotated variables lead to improved embryo selection. This thesis has explored the evidence that both supports or challenges the routine use of time-lapse in IVF laboratories. This constitutes an ongoing debate since there remains insufficient good quality evidence of difference for live birth or miscarriage with or without the use of time-lapse. Given the paucity of evidence, one cannot be absolutely certain, and conclusions drawn from many existing studies are low in quality. That said, the notion of improved selection through time-lapse continues to be challenged as more recent, well designed randomised controlled trials have found no difference in clinical outcomes with the use of time-lapse. This thesis has, however, recognised the diversity of morphokinetic algorithms available and therefore the intervention across these studies is inconsistent. Arguably therefore, this debate is not yet settled, nevertheless it has provided the context for a more specific question concerning the application of morphokinetics to determine risk of aneuploidy. This is a timely and relevant topic as there is now an appetite for non-invasively detecting the ploidy status of pre-implantation embryos in ART.

The debate of embryo selection isn't limited to morphokinetics but also PGT-A. This thesis has critically discussed available evidence from cohort studies and randomised controlled trials. Namely, it has been suggested that a more appropriate way to assess this

technology would be to report live birth rate per embryo transfer, include data on time to pregnancy and conduct further trials on poorer prognosis groups in order to address the limitations of existing studies. These are the important considerations for patients. Re-analysis of existing datasets in this way has highlighted the positive impact PGT-A can potentially have. That said, PGT-A has cost implications, technical challenges, a limited scope of genetic screening, carries a small risk of damage and many patients have ethical concerns. Therefore, it seems fitting that we endeavour to find a non-invasive alternative. To that end, there have been a variety of attempts to utilise morphokinetics in order to create an algorithm with the ability of risk stratifying embryos for ploidy status.

In order to explore this further, this thesis first explored the unadjusted association of morphokinetic variables and ploidy status through systematic review and meta-analysis. Interestingly, ten morphokinetic variables were delayed in aneuploid embryos. This was most notable later in embryonic development, around the time of blastulation, for example, t8 had a mean difference of 1.13h (95% CI 021-2.05; three studies, n=742, I^2=0%). Importantly, however, there is a significant amount of variability in the morphokinetic timings of euploid and aneuploid embryos; with significant heterogeneity in some variables. For example, while tB had a larger mean difference between euploid and aneuploid embryos of 2 hours, the confidence intervals were wide and close to zero (95% CI 0.15-3.81) with an I^2 of 76%. This chapter concludes that while utilising morphokinetics to accurately determine ploidy status will be impossible, it may be possible to utilise time-lapse to determine risk of aneuploidy through sophisticated algorithms. Importantly, morphokinetics have been found to be related to other factors,

including patient specific characteristics such as age and laboratory conditions. These confounding factors should therefore be considered when interpreting and planning future studies.

A variety of morphological aspects and their association with ploidy status were also explored in this thesis. These factors can be witnessed on a time-lapse device, thus if identified to be a potential prognostic marker, they could be incorporated into a future prediction model. Four additional literature searches were conducted for fragmentation, multinucleation, abnormal cleavage and embryo contraction. It was reported in this chapter that there is potentially some prognostic information in only: percentage of fragmentation, multinucleation persisting to the four-cell stage and frequency of embryo contractions. This was, however, very low-quality evidence and therefore we cannot be certain whether these morphological components are useful for incorporation into a ploidy prediction model.

The following chapter of this thesis explores different approaches to the development of morphokinetic models to predict ploidy status. Importantly, newer approaches using artificial intelligence were compared to more traditional statistical methods. The application of artificial intelligence in reproductive medicine was discussed, in particular it's growing popularity for embryo selection. Limitations were highlighted, including the interpretability of these algorithms, reduced effectiveness where only limited training data exists, and finally overfitting, thus potentially reducing external validity. This chapter compares twelve machine learning models developed to predict ploidy using a dataset of

8147 biopsied embryos. Four algorithms were used to build twelve models, mixed effects multivariable logistic regression, extreme gradient boosting, random forest classifier and deep learning. Interestingly, logistic regression outperformed all other approaches for discriminating between euploid and aneuploid embryos (AUC=0.71). Further questioning the increasing focus on artificial intelligence methodologies; such approaches may not be the most appropriate for more modest numerical datasets.

Following the identification of an appropriate methodology for model development, the work of this thesis then moved on to discuss the association of the risk scores derived from this model with live birth and miscarriage. This chapter also importantly investigates whether it's appropriate to include demographic or clinical parameters into embryo selection models. It was discovered that by including variables such as age, results in little ranking within a patient's cohort of embryos. We know that age is the best predictor of success and chance of aneuploidy, therefore the algorithms apply too much weight to this predictor, such that it ineffectively ranks at the level of the embryo. It was therefore concluded that such factors should not be included into morphokinetic models. Considering this, the 'morphokinetics only' logistic regression model had an AUC of 0.61 for predicting ploidy status. A separate dataset of 3587 embryos was used to test the association of this model's risk scores with live birth and miscarriage. An embryo deemed 'low risk' of aneuploidy as determined by this model was significantly more likely to result in a livebirth than those embryos graded 'high risk' (OR, 1.95; 95% CI, 1.65-2.25). In contrast, miscarriage was not associated with this model's risk score between the 'high risk' to 'moderate risk' or 'high risk' to 'low risk' embryos.

The final study in this thesis investigated whether morphokinetic models are better at prioritising euploid embryos for transfer over morphological selection by an embryologist. Data from this study were obtained from 1958 biopsied blastocysts. Two models were applied, a live birth and ploidy prediction model. It was identified that the use of the live birth and ploidy prediction morphokinetic models improved the prioritisation of a euploid embryo from 39% to 46% and 47%, respectively. A statistic termed normalised discounted cumulative gain (NDCG) was used, this is commonly used in machine learning research to determine the ranking ability of search engines or algorithms. A NDCG score of 1 indicates perfect ranking. There was a significant difference between the NDGC metric of the PREFER model vs. embryologist selection at 0.96 and 0.87, respectively ($t=14.1$, $p<0.001$). This study therefore demonstrated that there was improved prioritisation of euploid embryos with the application of a morphokinetic model.

This body of work has had several strengths associated, firstly, we have presented the first systematic review investigating associations between individual prognostic variables identified on a TLS. This is thus far the most comprehensive piece of work published on this topic, involving 58 studies. Second, the cohort studies reported in this thesis have included a large number of embryos and the models developed used robust methodology designed with experts in the field. Furthermore, the prognostic models have each been extensively externally validated using a variety of performance metrics. The weaknesses with the studies reported above are the fact that they are retrospective in design; while showing important associations with clinical outcomes, some performance metrics, e.g.

AUC 0.61, may be interpreted as a poorly performing model. This identifies the need to add further prognostic information into future models, such as image analysis.

Through a series of cohort studies and meta-analysis, we now have excellent grounds to justify a multi-arm randomised controlled trial (RCT). This would be the first RCT investigating a morphokinetic algorithm predicting aneuploidy. It would be vital to incorporate a representative cohort of patients, powered to allow for subgroup analyses. It is suggested that the outcome is live birth per embryo transfer and to include important secondary outcomes, such as time to pregnancy. This work has been a privilege to undertake, particularly since it has the potential to meaningful for patients.

APPENDIX I- SUPPLEMENTARY DATA FOR CHAPTER 2

Systematic Review MESH Search Terms

Embryo [AND] ploidy [AND] [[morphokinetics [OR] timelapse [OR] fragmentation [OR] abnormal cleavage [OR] reverse cleavage [OR] unequal cleavage [multinucleation]]

Supplementary Table I: QUIPS Risk of bias assessment instrument for prognostic factor studies

Modified from: Hayden JA, Côté P, Bombardier C. Evaluation of the Quality of Prognosis Studies in Systematic Reviews. Annals of Internal Medicine. 2006;144:427-437.

1. Study Participation	Goal: To judge the risk of selection bias (likelihood that relationship between morphokinetics and aneuploidy is different for participants and eligible non-participants).
Source of target population	The source population or population of interest is adequately described for key characteristics including age, BMI, reason for PGT-A, stimulation methods, biopsy technique
Method used to identify population	The sampling frame and recruitment are adequately described, including methods to identify the sample sufficient to limit potential bias
Recruitment period	Period of recruitment is adequately described
Place of recruitment	Place of recruitment (setting and geographic location) are adequately described
Inclusion and exclusion criteria	Inclusion and exclusion criteria are adequately described
Baseline characteristics	The baseline study sample (i.e., individuals entering the study) is adequately described for at the very least age, BMI, reason for PGT-A.
Summary Study participation	**The study sample represents the population of interest on key characteristics, sufficient to limit potential bias of the observed relationship between PF and outcome.**
2. Study Attrition	Goal: To judge the risk of attrition bias (likelihood that relationship between morphokinetics and aneuploidy are different for completing and non-completing participants).
Proportion of baseline sample available for analysis	The proportion of PGT-A embryos with valid and interpretable results
Reasons clearly explained for embryos which have no result after PGT-A	Does the study explain why some embryos are not included in the final analysis e.g. amplification failure
Potential impact of attrition	Is the reduced proportion of embryos with PGT-A results available likely to introduce bias?
Study Attrition Summary	**Loss to follow-up (from baseline sample to study population analysed) is not associated with key characteristics (i.e., the study data adequately represent the sample) sufficient to limit potential bias to the observed relationship between Morphokinetics and aneuploidy.**
3. Prognostic Factor Measurement	Goal: To judge the risk of measurement bias related to how morphokinetics was measured (differential measurement of morphokinetics related to the level of outcome).
Definition of Morphokinetics	A clear definition or description of the morphokinetic variables used including the thresholds for each parameter
Valid and Reliable Measurement of Morphokinetics	Method of morphokinetic measurement is validated across embryologists that annotate the embryos on the time lapse system- describes some methods of internal validation or inter-rater reliability via statistical analysis e.g. kappa statistics
Method and Setting of Morphokinetic Measurement	The method and setting of measurement of Morphokinetics is the same for all study participants.

Proportion of data on Morphokinetics available for analysis	Were all embryos that were included with PGT-A results cultured in the time lapse system and had morphokinetic analysis
Method used for missing data	Appropriate methods of imputation are used for missing morphokinetic data.
PF Measurement Summary	**Morphokinetics is adequately measured in study participants to sufficiently limit potential bias.**
4. Outcome Measurement	Goal: To judge the risk of bias related to the measurement of aneuploidy
Definition of the Outcome	A clear definition of aneuploidy is provided, for example in studies using next generation sequencing what levels of mosaicism do they categorise as aneuploidy
Valid and Reliable Measurement of Outcome	The method of assessing chromosomal status is validated
Method and Setting of Outcome Measurement	The method and setting of aneuploidy measurement is the same for all study participants.
Outcome Measurement Summary	**Aneuploidy is adequately measured in study participants to sufficiently limit potential bias.**
5. Study Confounding	Goal: To judge the risk of bias due to confounding (i.e. the effect of Morphokinetics is distorted by another factor that is related to Morphokinetics and Aneuploidy).
Important Confounders Measured	All important confounders, including age, BMI, stimulation drugs, reason for PGT-A are described
Definition of the confounding factor	Clear definitions of the important confounders measured are provided e.g. stimulation dosages used in aneuploidy vs. euploid groups
Method used for missing data	Appropriate methods are used if imputation is used for missing confounder data.
Appropriate Accounting for Confounding	Important potential confounders are accounted for in the study design (e.g., matching for key variables, stratification, or initial assembly of comparable groups).
	Important potential confounders are accounted for in the analysis (i.e., appropriate adjustment).
Study Confounding Summary	**Important potential confounders are appropriately accounted for, limiting potential bias with respect to the relationship between Morphokinetics and outcome.**
6. Statistical Analysis and Reporting	Goal: To judge the risk of bias related to the statistical analysis and presentation of results.
Presentation of analytical strategy	There is sufficient presentation of data to assess the adequacy of the analysis.
Model development strategy	The strategy for model building (i.e., inclusion of variables in the statistical model) is appropriate and is based on a conceptual framework or model.
	The selected statistical model is adequate for the design of the study.
Reporting of results	There is no selective reporting of results.
Statistical Analysis and Presentation Summary	**The statistical analysis is appropriate for the design of the study, limiting potential for presentation of invalid or spurious results.**

**Morphokinetics' was substituted for each prognostic factor tested by each separate literature search in this review*

Key to QUIPS assessment tool

Each of the six bias domains were rated for their overall risk of bias after considering each prompting item for that domain (Hayden et al)

Global bias rating	Study Participation	Study Attrition*	Prognostic factor (PF) measurement***	Outcome measurement**	Study Confounding	Statistical Analysis
High risk of bias	The relationship between the PF and outcome is very likely to be different for participants and eligible non-participants	The relationship between the PF and outcome is very likely to be different for competing and noncompeting participants	The measurement of the PF is very likely to be different for different levels of the outcome of interest	The measurement of the outcome is very likely to be different related to the baseline level of the prognostic factor	The observed effect of the PF on the outcome is very likely to be distorted by another factor related to PF and outcome	The reported results are very likely to be spurious or biased related to analysis or reporting
Moderate risk of bias	The relationship between the PF and outcome may be different for participants and eligible non-participants	The relationship between the PF and outcome may be different for competing and noncompeting participants	The measurement of the PF may be different for different levels of the outcome of interest	The measurement of the outcome may be different related to the baseline level of the prognostic factor	The observed effect of the PF on the outcome may be distorted by another factor related to PF and outcome	The reported results may be spurious or biased related to analysis or reporting
Low risk of bias	The relationship between the PF and the outcome is unlikely to be different for participants and eligible nonparticipants	The relationship between the PF and outcome is unlikely to be different for competing and noncompeting participants	The measurement of the PF is unlikely to be different for different levels of the outcome of interest	The measurement of the outcome us unlikely to be different related to the baseline level of the prognostic factor	The observed effect of the PF on the outcome is unlikely to be distorted by another factor related to PF and outcome	The reported results are unlikely to be spurious or biased related to analysis or reporting

Supplementary Table II: Characteristics of the excluded studies

Prognostic factor search	Author, year	Title	Reason excluded
Morphokinetics	Amir et al., 2019	Time-lapse imaging reveals delayed development of embryos carrying unbalanced chromosomal translocations.	Measured different outcome to aneuploidy e.g., translocations
Morphokinetics & Multinucleation	Lammers et al., 2019	Morphokinetic parameters in chromosomal translocation carriers undergoing preimplantation genetic testing.	Measured different outcome to aneuploidy e.g., translocations
Morphokinetics	Pons et al., 2019	Deconstructing the myth of poor prognosis for fast-cleaving embryos on day 3. Is it time to change the consensus?.	Measured developmental competence
Morphokinetics	Mumusoglu et al., 2017	Duration of blastulation may be associated with ongoing pregnancy rate in single euploid blastocyst transfer cycles.	Clinical outcomes only
Morphokinetics	Rocafort et al., 2016	Euploid embryos selected by an automated time-lapse system have superior SET outcomes than selected solely by conventional morphology assessment.	Investigates a model to identify euploid embryos with the best clinical outcomes but doesn't test its ability to discriminate ploidy
Morphokinetics	Balakier et al ., 2016	Impact of multinucleated blastomeres on embryo developmental competence, morphokinetics, and aneuploidy.	Focuses on subset of embryos e.g., multinucleated
Morphokinetics	Hashimoto et al. 2016	Impact of multinucleated blastomeres on embryo developmental competence, morphokinetics, and aneuploidy.	Focuses on subset of embryos e.g., multinucleated
Morphokinetics	Bayram et al., 2019	Cleavage stage mitochondrial DNA is correlated with preimplantation human embryo development and ploidy status.	Focuses on subset of embryos e.g., multinucleated
Morphokinetics	Escriba et al., 2016	Kinetics of the early development of uniparental human haploid embryos.	Focuses on subset of embryos e.g., multinucleated
Morphokinetics	Mateo et al., 2017	Could monopronucleated ICSI zygotes be considered for transfer? Analysis through time-lapse monitoring and PGS.	Focuses on subset of embryos e.g., multinucleated
Morphokinetics	Grau et al., 2015	Morphokinetics as a predictor of self-correction to diploidy in tripronucleated intracytoplasmic sperm injection-derived human embryos.	Focuses on subset of embryos e.g., multinucleated
Morphokinetics	Kahraman et al. 2020	High rates of aneuploidy, mosaicism and abnormal morphokinetic development in cases with low sperm concentration.	Inappropriate sample population
Morphokinetics	Bar-El et al. 2016	Blastomere biopsy for PGD delays embryo compaction and blastulation: a time-lapse microscopic analysis.	PGT-A data not reported as an outcome
Morphokinetics	Gazzo et al. 2020	The KidscoreTM D5 algorithm as an additional tool to morphological assessment and PGT-A in embryo selection: a time-lapse study.	Clinical outcomes only & no morphokinetic data
Morphokinetics	Chavez-Badiola et al. 2020	Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation.	Different prognostic factor tested
Morphokinetics	De Munck et al 2021	Intracytoplasmic sperm injection is not superior to conventional IVF in couples with non-male factor infertility and preimplantation genetic testing for aneuploidies (PGT-A).	Different prognostic factor tested

Morphokinetics	Derrick et al., 2017	Perivitelline threads associated with fragments in human cleavage stage embryos observed through time-lapse microscopy.	Different prognostic factor tested
Morphokinetics	Gazzo et al., 2020	Blastocyst contractions are strongly related with aneuploidy, lower implantation rates, and slow-cleaving embryos: a time lapse study.	Different prognostic factor tested
Morphokinetics	Huang et al., 2019	Early blastocyst expansion in euploid and aneuploid human embryos: evidence for a non-invasive and quantitative marker for embryo selection.	Different prognostic factor tested
Morphokinetics	Gonzalex et al., 2018	Contraction behaviour reduces embryo competence in high-quality euploid blastocysts.	Different prognostic factor tested
Morphokinetics	Lagalla et al., 2017	Embryos with morphokinetic abnormalities may develop into euploid blastocysts.	Different prognostic factor tested
Morphokinetics	McCoy et al., 2018	Tripolar chromosome segregation drives the association between maternal genotype at variants spanning PLK4 and aneuploidy in human preimplantation embryos.	Different prognostic factor tested
Morphokinetics	Ottolini et al., 2017	Tripolar mitosis and partitioning of the genome arrests human preimplantation development in vitro.	Different prognostic factor tested
Morphokinetics	Ozbek et al., 2021	Comparison of single euploid blastocyst transfer cycle outcome derived from embryos with normal or abnormal cleavage patterns.	Different prognostic factor tested
Morphokinetics	Lagalla et al., 2020	Alternative patterns of partial embryo compaction: prevalence, morphokinetic history and possible implications.	Different prognostic factor tested
Morphokinetics	Vera-Rodriguez et al., 2015	Prediction model for aneuploidy in early human embryo development revealed by single-cell analysis.	Morphokinetic measurement not in a standardised format
Morphokinetics	Schenk et al., 2018.	Impact of polar body biopsy on embryo morphokinetics-back to the roots in preimplantation genetic testing?.	Polar body biopsy
Morphokinetics	Campbell et al., 2014	Aneuploidy is a key causal factor of delays in blastulation: author response to 'A cautionary note against aneuploidy risk assessment using time-lapse imaging'.	Commentary
Fragmentation	Stone et al., 2019	Embryo fragmentation as a determinant of blastocyst development in vitro and pregnancy outcomes following embryo transfer.	Investigated blastocyst development competence in relation to fragmentation only
Fragmentation	Ebner et al., 2003	Selection based on morphological assessment of oocytes and embryos at different stages of preimplantation development: a review.	Literature Review
Fragmentation	Pellestor et al. 1995	Relationship between morphology and chromosomal constitution in human preimplantation embryo.	Literature Review
Fragmentation	Daughtry et al., 2019	Single-cell sequencing of primate preimplantation embryos reveals chromosome elimination via cellular fragmentation and blastomere exclusion.	Ploidy status not the outcome measured
Fragmentation	Bongso et al., 1991	Preimplantation genetics: chromosomes of fragmented human preembryos.	Ploidy status not the outcome measured
Fragmentation & Multinucleation	Munne et al., 1998	Chromosome abnormalities in human embryos.	Ploidy status not the outcome measured- only mosaicism
Fragmentation	Pellestor et al., 1994	Relationship between morphology and chromosomal constitution in human preimplantation embryo.	Investigated morphology grading not specifically fragmentation

Fragmentation	Almeida et al., 1994	The relationship between chromosomal abnormalities in the human oocyte and fertilization in vitro, Human Reproduction	Examined oocytes not embryos
Abnormal Cleavage	Nagai et al., 2021	Abnormal cleavage is involved in the self-correction of bovine preimplantation embryos.	Non-human embryos
Abnormal Cleavage	Magata et al., 2019	Growth potential of bovine embryos presenting abnormal cleavage observed through time lapse cinematography.	Non-human embryos
Abnormal Cleavage	Okada et al. 2020	Analysis of chromosomal abnormality of bovine IVF embryos based on next-generation sequencing.	Non-human embryos
Abnormal Cleavage	Han et al., 1999	Pronuclear location before the first cell division determines ploidy of polyspermic pig embryos.	Different prognostic factor tested
Abnormal Cleavage	Boediono et al., 2021	Morphokinetics of embryos after IMSI versus ICSI in couples with sub-optimal sperm quality: A time-lapse study.	Different prognostic factor tested
Abnormal Cleavage	Lagalla et al. 2020	Alternative patterns of partial embryo compaction: prevalence, morphokinetic history and possible implications.	Different prognostic factor tested
Abnormal Cleavage	Davies et al., 2018	Male Factor is the most important factor influencing the frequency of unequal direct cleavage events as visualized by time-lapse during early embryo development.	Ploidy status not the outcome measured
Abnormal Cleavage	Montgomery et al. 2018	Time Lapse assessment of the occurrence and clinical outcome of direct cleavage in a population of 10,529 embryos cultured to the blastocyst stage.	Clinical outcomes only
Abnormal Cleavage	Cetinkaya et al., 2014	The synchronicity of mitotic divisions predicts embryo implantation and live birth.	Clinical outcomes only
Abnormal Cleavage	Karamalegos et al. 2014	Assessing the impact of factors derived from morphokinetic analysis in decreasing implantation potential of morphologically good embryos.	Clinical outcomes only
Abnormal Cleavage	Davies et al., 2016	Time Lapse analysis of the interrelationship between direct cleavage, multinucleation and maternal age in Natural Cycle IVF and Standard IVF.	Ploidy status not the outcome measured
Abnormal Cleavage	Grau et al. 2015	Morphokinetics as a predictor of self-correction to diploidy in tripronucleated intracytoplasmic sperm injection-derived human embryos.	Focuses on subset of embryos e.g., multinucleated
Abnormal Cleavage	Burruel et al., 2014	Abnormal early cleavage events predict early embryo demise: sperm oxidative stress and early abnormal cleavage.	Ploidy status not the outcome measured
Abnormal Cleavage	Alfarawati et al., 2012	How does aneuploidy affect embryo morphology and development from the cleavage to the blastocyst stage.	Investigated morphology grading not specifically fragmentation
Abnormal Cleavage	Ho et al., 2017	Abnormal cleavage patterns in embryos are associated with aneuploidy and poor morphology scores.	Duplicate abstract of the original paper by Ho et al. 2018
Abnormal Cleavage and Morphokinetics	Ho et al., 2017	Blastulation timing is associated with differential mitochondrial content in euploid embryos	Morphokinetic measurement not in a standardised format
Contraction	Chian et al., 2016	Comparing blastocyst expansion dynamics between euploid vs. Aneuploid embryos: A quantitative and automated analysis of time-lapse cinematography.	Investigated association of blastocyst expansion volume/ rate of expansion and ploidy
Contraction	Huang et al., 2017	Comparison of blastocyst expansion morphokinetics in euploid versus aneuploid embryos from infertility patients.	Investigated association of blastocyst expansion volume/ rate of expansion and ploidy

Contraction	Huang et al., 2017	Morphokinetics of blastocyst expansion in euploid and aneuploid human embryos.	Investigated association of blastocyst expansion volume/ rate of expansion and ploidy
Contraction	Huang et al., 2019	Early blastocyst expansion in euploid and aneuploid human embryos: evidence for a non-invasive and quantitative marker for embryo selection.	Investigated association of blastocyst expansion volume/ rate of expansion and ploidy
Contraction	Caprell et al., 2019	Increased expansion and decreased contraction of embryos corresponds to increased clinical pregnancy rates in single FET cycles.	Clinical outcomes only
Multinucleation	Laverge et al., 2000	Chromosome analysis of human preimplantation embryos by fluorescent in situ hybridization.	Review article
Multinucleation	Zhang et al. 2012	Clinical relevance of embryos with multinucleated blastomeres in PGD cycles for aneuploidy, translocation, or single gene defects.	Clinical outcomes only
Multinucleation	Edgar et al. 2012	Factors associated with multinucleation in human cleavage stage embryos.	Clinical outcomes only
Multinucleation	Karamalegos et al., 2014	Assessing the impact of factors derived from morphokinetic analysis in decreasing implantation potential of morphologically good embryos.	Clinical outcomes only
Multinucleation	Zhan et al., 2018	Detection of multinucleated trophectoderm cells by time-lapse microscopy: Implications for PGT-A biopsy results.	Investigates multinucleated cells in the trophectoderm
Multinucleation	Munne, 1993	Unsuitability of multinucleated human blastomeres for preimplantation genetic diagnosis.	Focuses on sex chromosomes only to determine origins or multinucleation

Supplementary Table III: Summary of correspondence to the date of publication

Author, year (see reference list for full details)	Contacted	Reason	Response	Decision
Kramer et al., 2014	Yes	Use time of syngamy rather than time of insemination as start point	Received and data given but not able to compare	Include but excluded from quantitative analysis
Ho et al., 2018	Yes	Used pronuclear breakdown as start point rather than time of insemination	No response	Exclude
Del Carmen Nogales et al., 2017	Yes	Separates morphokinetics for euploid, monosomy, trisomy and complex aneuploidy	No response	Include but excluded from quantitative analysis
Chavez et al. 2012	Yes	Some morphokinetic variables not reported	No response	Include
Kimelman et al., 2019	Yes	No standard deviations/ measure of dispersion displayed	No response	Include
Moayeri et al. 2008	No- email undeliverable and no alternatives	Contacted for further data/ supplementary information	NA	Include but excluded from quantitative analysis
Chavez et al. 2012	Yes	Contacted for further data/ supplementary information	No response	Include
Vera-Rodriguez et al., 2015	Yes	Contacted for further data/ supplementary information	No response	Include but excluded from quantitative analysis
Magli et al., 2001	Yes	To confirm that both studies contain different embryos- confirmed different	Response partially received	Include. Unable to ascertain raw numbers from figure to extract data, contact unsuccessful for this.
Magli et al.,2007	Yes	As above	Response received	Include
Delimitreva et al. 2005	Yes	Contacted for further data/ supplementary information	No response	Include
Ziebe et al., 2003	Yes	Contacted for further data/ supplementary information	No response	Include
Rienzi et al., 2013	Yes	Contacted for further data/ supplementary information	No response	Include
Zhan et al., 2015	Yes	Contacted for further data/ supplementary information	No response	Include
Lagalla et al., 2017 & Lagalla et al., 2015	Yes	Contacted to confirm this study contains the same patients as a published abstract by the same authors (Lagalla et al. 2015)	No response	Original article included; abstract is 'awaiting classification' due to unanswered correspondence. Likely the same embryos using intuition.
McCoy et al., 2018	No- email undeliverable,	Contacted for further data/ supplementary information	NA	Include

	unable to find alternative			
Ozbek et al., 2021	Yes	Contacted for further data/ supplementary information	Responded	Include
Gonzalez et al., 2018	Yes	Contacted for further data/ supplementary information	No response	Include
Gazzo et al., 2019	Yes	Contacted for further data/ supplementary information	No response	Include
Scott et al., 2010	Yes	Contacted for further data/ supplementary information	No response	Include but excluded from quantitative analysis
Davies et al. 2016	No- email undeliverable, unable to find alternative	Contacted for further data/ supplementary information	No response	Include but excluded from quantitative analysis
Melzer et al., 2013	Yes	Contacted for further data/ supplementary information	No response	Include but excluded from quantitative analysis
Yilmaz et al., 2013	Yes	Contacted for further data/ supplementary information	No response	Include but excluded from quantitative analysis
Li et al., 2015	Yes	Contacted for further data/ supplementary information	No response	Include
Goodman et al., 2015	Yes	Contacted for further data/ supplementary information	No response	Include but excluded from quantitative analysis
Desai et al., 2016 & Desai et al., 2018	Yes	Contact to confirm abstract published is different to Desai et al. 2018 paper.	No response	Abstract remains in the 'awaiting classification table' and not included due to risk of it containing the same embryos. Desai et al. 2018 included.
Agerholm et al., 2008	Not able to find updated contact information	Contacted for further data/ supplementary information	No response	Include
Mazur et al. 2013	Yes	Contacted for further data/ supplementary information	Responded	Include

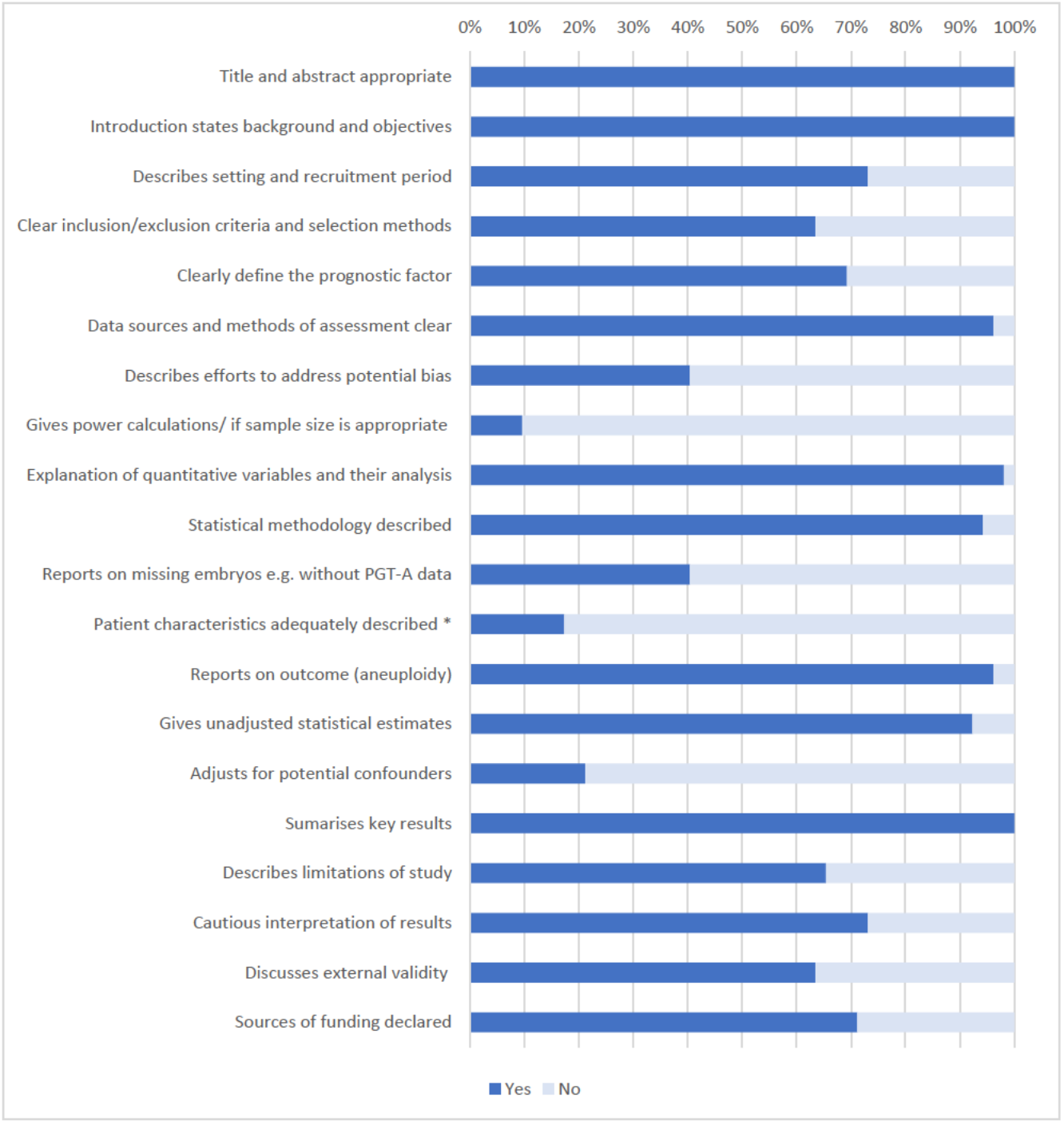
Supplementary Table 4: Assessment of quality of included studies using the adapted QUIPS tool described in Table 2

Study First Author, Year	Study Participation	Study Attrition	Prognostic Factor Measurement	Outcome measurement	Study Confounding	Statistical Analysis and Reporting
Morphokinetics and Ploidy						
Chavez, 2013	Moderate	Low	Moderate	Low	Moderate	Moderate
Campbell, 2013a	Moderate	Moderate	Low	Low	Moderate	Low
Campbell, 2013b	Moderate	Moderate	Low	Moderate	Low	Moderate
Yang, 2014	Low	Low	Low	Low	Moderate	Low
Basile, 2014	Low	Moderate	Low	Low	Moderate	Low
Kramer, 2014	Moderate	Low	Low	Low	Moderate	Low
Chawla, 2015	Moderate	High	Low	Low	Moderate	Moderate
Patel, 2016	Low	Moderate	Low	Low	Moderate	Moderate
Minasi, 2016	Moderate	Low	Low	Low	Moderate	Low
Mumusoglu, 2016	Low	Low	Low	Low	Low	Low
Del Carmen Nogales, 2017	Moderate	Low	Low	Low	Moderate	Moderate
Rienzi, 2015	Moderate	Moderate	Low	Low	Moderate	Low
Zhang, 2017	Moderate	Moderate	Low	Low	Moderate	Low
Desai, 2018	Low	Low	Low	Low	Moderate	Moderate
Lee, 2018	Low	Low	Low	Low	Moderate	Low
Kimelman, 2019	Low	Low	Low	Low	Moderate	Moderate
Martin, 2021	Low	Low	Moderate	Low	Low	Low
Fragmentation and ploidy						
Magli, 2001	Moderate	Moderate	Moderate	High	Moderate	Low
Ziebe, 2003	Low	Low	Low	High	Low	Low
Delimitreva, 2005	Low	Low	Moderate	High	Moderate	Moderate
Magli, 2007	Moderate	Moderate	Moderate	High	Moderate	Low
Moayeri, 2008	Low	High	Moderate	High	Low	Moderate
Chavez, 2012	Low	Moderate	Low	Low	Moderate	Moderate

Vera-Rodriguez, 2015	Moderate	Moderate	Low	Low	Moderate	Moderate
Minasi, 2016	Moderate	Low	Moderate	Low	Low	Low
Abnormal cleavage and ploidy						
Campbell, 2013a	Moderate	Moderate	Low	Low	Moderate	Low
Rienzi, 2013	Moderate	Moderate	Low	Low	Moderate	Moderate
Davies, 2014	Moderate	Moderate	Moderate	Low	Moderate	High
Vera-Rodriguez, 2015	Moderate	Moderate	Low	Low	Moderate	Low
Zhan, 2016	Moderate	Moderate	Low	High	Moderate	Low
Lagalla, 2017	Moderate	Low	Low	Low	Moderate	Low
Zhang, 2017	Moderate	Moderate	Low	Low	Moderate	Low
Desai, 2018	Low	High	Low	Moderate	Moderate	Low
Ho, 2018	Moderate	Moderate	Low	Low	Low	Low
McCoy, 2018	Moderate	Moderate	Moderate	High	Moderate	Low
Ozbek, 2021	Low	Low	Low	Low	Low	Low
Contraction and ploidy						
Gonzalez, 2018	Moderate	Low	Low	Low	Low	Low
Gazzo, 2019	Moderate	Moderate	Low	Low	Low	Low
Multinucleation and ploidy						
Kligman, 1996	Moderate	High	Moderate	High	Moderate	Low
Magli, 2001	Moderate	Low	Moderate	High	Moderate	Low
Agerholm, 2008	Moderate	Low	Low	High	Moderate	Low
Scott, 2010	Moderate	Moderate	Moderate	Moderate	Low	Moderate
Ambroggio, 2011	Low	Low	Moderate	High	Low	Low
Davies, 2016	Moderate	High	Low	Low	Low	Moderate
Campbell, 2013	Moderate	Moderate	Low	Low	Moderate	Low
Mazur, 2013	Moderate	Moderate	Low	Low	Low	Low
Melzer, 2013	Moderate	Moderate	Low	Low	Moderate	Moderate
Munoz, 2014	Moderate	Moderate	Low	Moderate	Moderate	Low
Yilmaz, 2014	Moderate	High	Low	Low	Low	Moderate
Bayram, 2015	Moderate	Low	Low	Low	Moderate	Low

Li, 2015	Moderate	Moderate	Low	Low	Low	Moderate
Balakier, 2016	Moderate	Low	Low	Low	Moderate	Low
Goodman, 2015	Moderate	Low	Low	Moderate	Low	Low
Hashimoto, 2016	Low	High	Low	Low	Moderate	Low
Nogales, 2014	Moderate	Moderate	Low	Low	Moderate	Low
Zhang, 2017	Moderate	Moderate	Low	Low	Moderate	Low
Desai, 2018	Low	High	Low	Low	Moderate	Low
Lee, 2019	Low	Low	Low	Low	Moderate	Low

Supplementary Figure 1: STROBE Assessment of all papers meeting selection criteria completed according to published criteria (von Elm et al. 2007).



**Categorised as 'yes' if the study includes a demographics or patient characteristics tabl*

Supplementary Figures 2- Weighted mean difference and prognostic factor graphs for morphokinetic variables

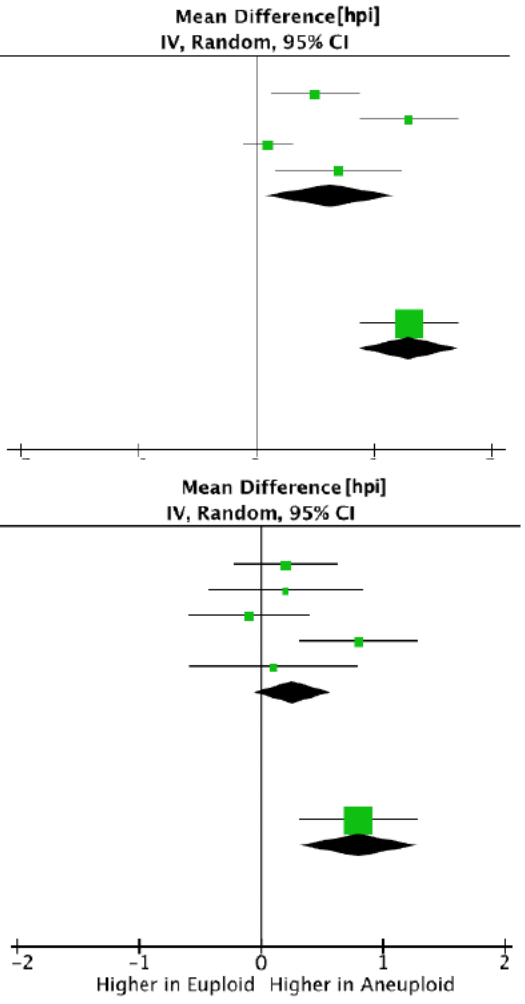
tPB2

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Chawla 2014	4.5	1.8	263	4	2.2	197	25.3%	0.50 [0.12, 0.88]	2014
Mumusoglu 2016	4.1	1.9	242	2.8	2.3	173	24.6%	1.30 [0.88, 1.72]	2016
Minasi 2016	4	1.9	634	3.9	1.3	297	27.6%	0.10 [-0.11, 0.31]	2016
Patel 2016	4	1.7	41	3.3	0.7	126	22.5%	0.70 [0.17, 1.23]	2016
Subtotal (95% CI)			**1180**			**793**	**100.0%**	**0.63 [0.09, 1.18]**	
Heterogeneity: $Tau^2 = 0.27$; $Chi^2 = 27.19$, df = 3 ($P < 0.00001$); $I^2 = 89\%$									
Test for overall effect: $Z = 2.27$ ($P = 0.02$)									
Excluded studies using FISH and D3/ blastomere biopsy									
Mumusoglu 2016	4.1	1.9	242	2.8	2.3	173	100.0%	1.30 [0.88, 1.72]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**1.30 [0.88, 1.72]**	
Heterogeneity: Not applicable									
Test for overall effect: $Z = 6.09$ ($P < 0.00001$)									

tPN

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Basile 2014	9	2.9	361	8.8	1.8	143	25.3%	0.20 [-0.22, 0.62]	2014
Chawla 2014	10.8	3.4	263	10.6	3.4	197	16.3%	0.20 [-0.43, 0.83]	2014
Minasi 2016	10.1	3.8	634	10.2	3.5	297	21.5%	-0.10 [-0.60, 0.40]	2016
Mumusoglu 2016	9.4	3	242	8.6	2	173	22.2%	0.80 [0.32, 1.28]	2016
Patel 2016	7.7	2.5	126	7.6	1.7	41	14.7%	0.10 [-0.58, 0.78]	2016
Subtotal (95% CI)			**1626**			**851**	**100.0%**	**0.25 [-0.06, 0.57]**	
Heterogeneity: $Tau^2 = 0.06$; $Chi^2 = 7.19$, df = 4 ($P = 0.13$); $I^2 = 44\%$									
Test for overall effect: $Z = 1.57$ ($P = 0.12$)									
Excluded studies using FISH and D3/ blastomere biopsy									
Mumusoglu 2016	9.4	3	242	8.6	2	173	100.0%	0.80 [0.32, 1.28]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**0.80 [0.32, 1.28]**	
Heterogeneity: Not applicable									
Test for overall effect: $Z = 3.26$ ($P = 0.001$)									

Test for subgroup differences: $Chi^2 = 3.45$, df = 1 ($P = 0.06$), $I^2 = 71.1\%$



tPNf

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Chawla 2014	25.8	5.6	263	24.5	4.3	197	16.3%	1.30 [0.40, 2.20]	2014
Minasi 2016	24.4	3.5	634	24.8	4.5	297	23.1%	-0.40 [-0.98, 0.18]	2016
Mumusoglu 2016	24.8	3.1	242	24.5	2.8	173	23.3%	0.30 [-0.27, 0.87]	2016
Patel 2016	7.7	2.5	126	7.7	2.5	41	16.7%	0.00 [-0.88, 0.88]	2016
Zhang 2017	22.5	3	106	22	2.4	150	20.6%	0.50 [-0.19, 1.19]	2017
Subtotal (95% CI)			**1371**			**858**	**100.0%**	**0.29 [-0.23, 0.81]**	

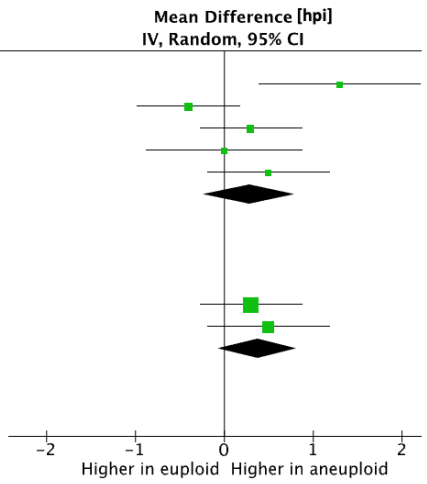
Heterogeneity: $Tau^2 = 0.22$; $Chi^2 = 10.82$, df = 4 (P = 0.03); $I^2 = 63\%$
Test for overall effect: Z = 1.11 (P = 0.27)

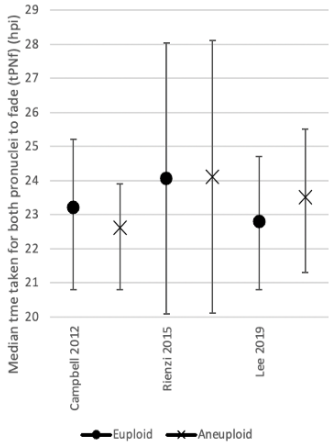
Excluded studies using FISH and D3/ blastomere biopsy

Mumusoglu 2016	24.8	3.1	242	24.5	2.8	173	59.2%	0.30 [-0.27, 0.87]	2016
Zhang 2017	22.5	3	106	22	2.4	150	40.8%	0.50 [-0.19, 1.19]	2017
Subtotal (95% CI)			**348**			**323**	**100.0%**	**0.38 [-0.06, 0.82]**	

Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.19$, df = 1 (P = 0.66); $I^2 = 0\%$
Test for overall effect: Z = 1.70 (P = 0.09)

Test for subgroup differences: $Chi^2 = 0.07$, df = 1 (P = 0.80), $I^2 = 0\%$





t2

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Chawla 2014	30.6	9.7	263	28.3	7.2	197	3.1%	2.30 [0.76, 3.84]	2014
Basile 2014	26.8	3.4	361	26.8	0.7	143	22.7%	0.00 [-0.37, 0.37]	2014
Minasi 2016	28.4	0.9	634	28	4.4	297	17.0%	0.40 [-0.11, 0.91]	2016
Mumusoglu 2016	27.3	3.3	242	26.7	2.9	173	14.0%	0.60 [0.00, 1.20]	2016
Patel 2016	27.6	3.5	126	27.6	3	41	5.6%	0.00 [-1.10, 1.10]	2016
Zhang 2017	25	3	106	24.5	2.5	150	11.5%	0.50 [-0.20, 1.20]	2017
Desai 2018	26.2	0.3	394	25.9	2.73	319	26.0%	0.30 [-0.00, 0.60]	2018
Subtotal (95% CI)			**2126**			**1320**	**100.0%**	**0.36 [0.07, 0.64]**	

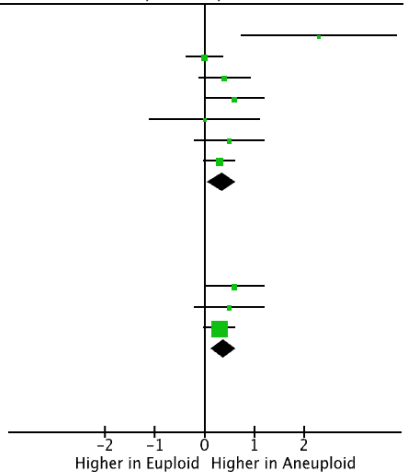
Heterogeneity: $Tau^2 = 0.06$; $Chi^2 = 10.69$, df = 6 (P = 0.10); $I^2 = 44\%$
Test for overall effect: Z = 2.47 (P = 0.01)

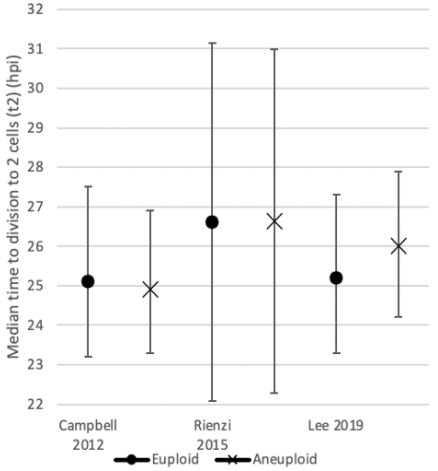
Excluded studies using FISH and D3/ blastomere biopsy

Mumusoglu 2016	27.3	3.3	242	26.7	2.9	173	17.5%	0.60 [0.00, 1.20]	2016
Zhang 2017	25	3	106	24.5	2.5	150	13.0%	0.50 [-0.20, 1.20]	2017
Desai 2018	26.2	0.3	394	25.9	2.73	319	69.5%	0.30 [-0.00, 0.60]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.38 [0.13, 0.63]**	

Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.90$, df = 2 (P = 0.64); $I^2 = 0\%$
Test for overall effect: Z = 2.96 (P = 0.003)

Test for subgroup differences: $Chi^2 = 0.01$, df = 1 (P = 0.92), $I^2 = 0\%$

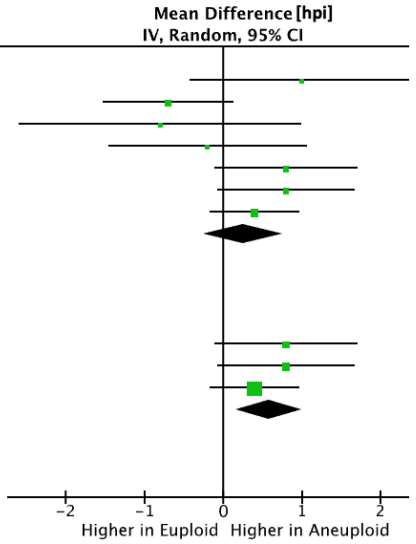


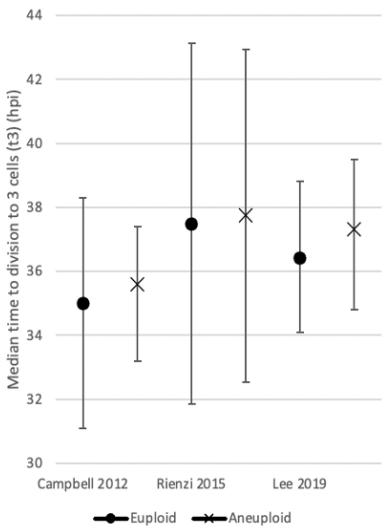


t3

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Chawla 2014	39.7	8.5	263	38.7	7	197	9.1%	1.00 [-0.42, 2.42]	2014
Basile 2014	37.2	4.8	361	37.9	3.96	143	17.6%	-0.70 [-1.52, 0.12]	2014
Patel 2016	37.9	5.2	41	38.7	4.6	126	6.4%	-0.80 [-2.58, 0.98]	2016
Minasi 2016	37.2	8.99	297	37.4	9.23	634	10.9%	-0.20 [-1.45, 1.05]	2016
Mumusoglu 2016	37.6	4.5	242	36.8	4.7	173	16.0%	0.80 [-0.10, 1.70]	2016
Zhang 2017	36	3.6	106	35.2	3.3	150	16.7%	0.80 [-0.07, 1.67]	2017
Desai 2018	37.1	4	394	36.7	3.6	319	23.3%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**1704**			**1742**	**100.0%**	**0.25 [-0.26, 0.75]**	
Heterogeneity: $Tau^2 = 0.20$; $Chi^2 = 11.35$, df = 6 (P = 0.08); $I^2 = 47\%$									
Test for overall effect: Z = 0.97 (P = 0.33)									
Excluded studies using FISH and D3/ blastomere biopsy									
Mumusoglu 2016	37.6	4.5	242	36.8	4.7	173	21.3%	0.80 [-0.10, 1.70]	2016
Zhang 2017	36	3.6	106	35.2	3.3	150	23.1%	0.80 [-0.07, 1.67]	2017
Desai 2018	37.1	4	394	36.7	3.6	319	55.5%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.58 [0.16, 0.99]**	
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.88$, df = 2 (P = 0.65); $I^2 = 0\%$									
Test for overall effect: Z = 2.72 (P = 0.007)									

Test for subgroup differences: $Chi^2 = 0.97$, df = 1 (P = 0.32), $I^2 = 0\%$

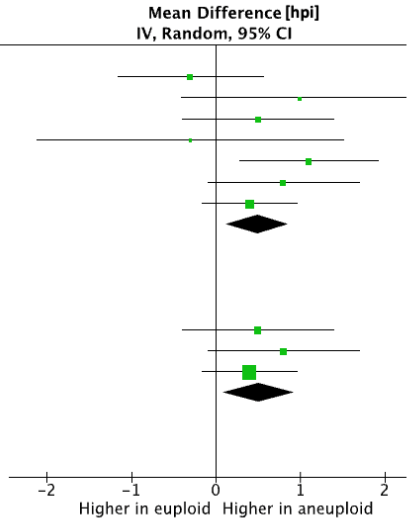


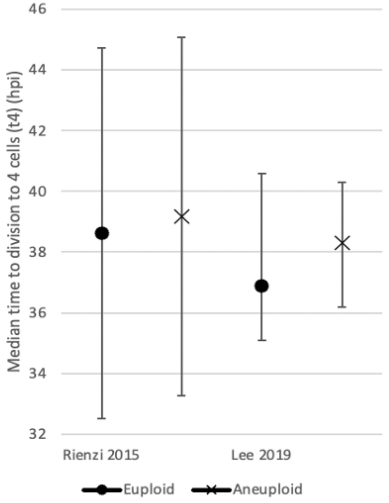


t4

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Basile 2014	39.6	4.84	361	39.9	4.27	143	15.3%	-0.30 [-1.16, 0.56]	2014
Chawla 2014	41.5	8.2	263	40.5	7.2	197	6.4%	1.00 [-0.41, 2.41]	2014
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	14.3%	0.50 [-0.39, 1.39]	2016
Patel 2016	40.1	5	126	40.4	5.2	41	4.0%	-0.30 [-2.12, 1.52]	2016
Minasi 2016	41.1	6.4	634	40	5.7	297	16.6%	1.10 [0.28, 1.92]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	14.4%	0.80 [-0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	29.0%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**2126**			**1320**	**100.0%**	**0.49 [0.12, 0.86]**	
Heterogeneity: $Tau^2 = 0.04$; $Chi^2 = 7.17$, df = 6 (P = 0.31); $I^2 = 16\%$									
Test for overall effect: Z = 2.60 (P = 0.009)									
Excluded studies using FISH and D3/ blastomere biopsy									
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	22.0%	0.50 [-0.39, 1.39]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	22.1%	0.80 [-0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	55.9%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.51 [0.09, 0.93]**	
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.55$, df = 2 (P = 0.76); $I^2 = 0\%$									
Test for overall effect: Z = 2.38 (P = 0.02)									

Test for subgroup differences: $Chi^2 = 0.00$, df = 1 (P = 0.95), $I^2 = 0\%$

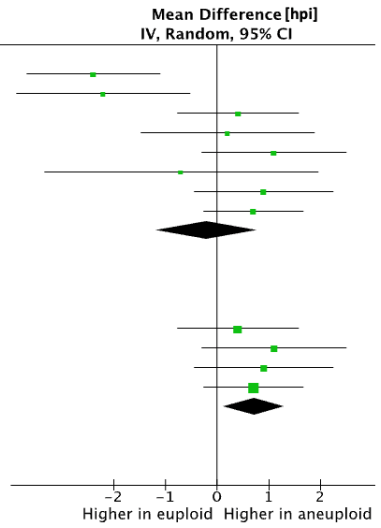


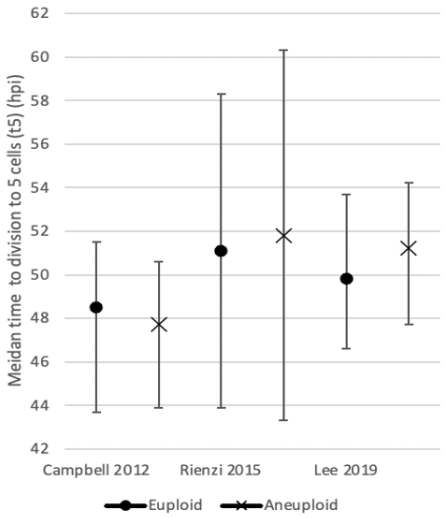


t5

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Basile 2014	49.4	7.75	361	51.8	6.1	143	13.5%	−2.40 [−3.68, −1.12]	2014
Chawla 2014	50.1	9.6	263	52.3	8.6	197	11.7%	−2.20 [−3.87, −0.53]	2014
Yang 2014	50.5	4.7	135	50.1	4.8	121	14.1%	0.40 [−0.77, 1.57]	2014
Minasi 2016	50.6	12.8	634	50.4	11.8	297	11.6%	0.20 [−1.47, 1.87]	2016
Mumusoglu 2016	49.7	7.2	242	48.6	7.1	173	13.0%	1.10 [−0.29, 2.49]	2016
Patel 2016	51.3	8.3	126	52	7.2	41	7.7%	−0.70 [−3.34, 1.94]	2016
Zhang 2017	49.2	5.3	106	48.3	5.5	150	13.3%	0.90 [−0.44, 2.24]	2017
Desai 2018	49.9	6.6	319	49.2	6.4	394	15.1%	0.70 [−0.26, 1.66]	2018
Subtotal (95% CI)			**2186**			**1516**	**100.0%**	**−0.19 [−1.17, 0.79]**	
Heterogeneity: $Tau^2 = 1.42$; $Chi^2 = 27.09$, df = 7 (P = 0.0003); $I^2 = 74\%$									
Test for overall effect: Z = 0.38 (P = 0.71)									
Excluded studies using FISH and D3/ blastomere biopsy									
Yang 2014	50.5	4.7	135	50.1	4.8	121	25.4%	0.40 [−0.77, 1.57]	2014
Mumusoglu 2016	49.7	7.2	242	48.6	7.1	173	17.8%	1.10 [−0.29, 2.49]	2016
Zhang 2017	49.2	5.3	106	48.3	5.5	150	19.3%	0.90 [−0.44, 2.24]	2017
Desai 2018	49.9	6.6	319	49.2	6.4	394	37.4%	0.70 [−0.26, 1.66]	2018
Subtotal (95% CI)			**802**			**838**	**100.0%**	**0.73 [0.15, 1.32]**	
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.64$, df = 3 (P = 0.89); $I^2 = 0\%$									
Test for overall effect: Z = 2.44 (P = 0.01)									

Test for subgroup differences: $Chi^2 = 2.50$, df = 1 (P = 0.11), $I^2 = 60.0\%$

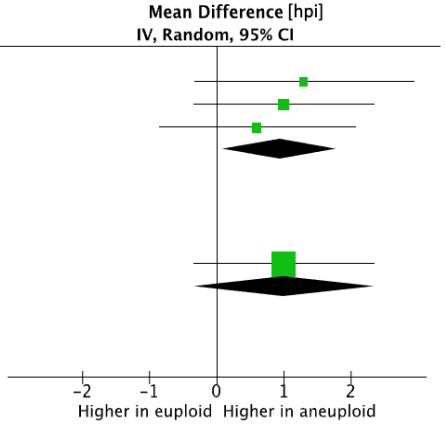


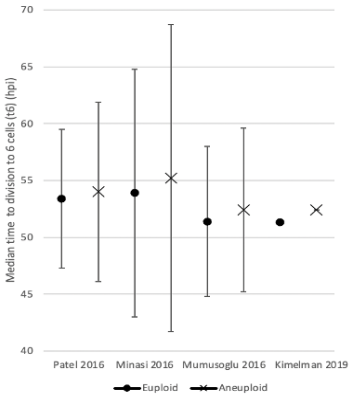


t6

Study or Subgroup	Aneuploidy Mean [hpi]	Aneuploidy SD [hpi]	Aneuploidy Total	Euploidy Mean [hpi]	Euploidy SD [hpi]	Euploidy Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Minasi 2016	55.2	13.5	634	53.9	11	297	26.7%	1.30 [−0.33, 2.93]	2016
Mumusoglu 2016	52.4	7.2	242	51.4	6.6	173	39.8%	1.00 [−0.34, 2.34]	2016
Patel 2016	54	7.9	126	53.4	6.1	634	33.5%	0.60 [−0.86, 2.06]	2016
Subtotal (95% CI)			**1002**			**1104**	**100.0%**	**0.95 [0.10, 1.79]**	
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.40$, df = 2 (P = 0.82); $I^2 = 0\%$									
Test for overall effect: Z = 2.20 (P = 0.03)									
Excluded studies using FISH and D3/ blastomere biopsy									
Mumusoglu 2016	52.4	7.2	242	51.4	6.6	173	100.0%	1.00 [−0.34, 2.34]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**1.00 [−0.34, 2.34]**	
Heterogeneity: Not applicable									
Test for overall effect: Z = 1.46 (P = 0.14)									

Test for subgroup differences: $Chi^2 = 0.00$, df = 1 (P = 0.95), $I^2 = 0\%$

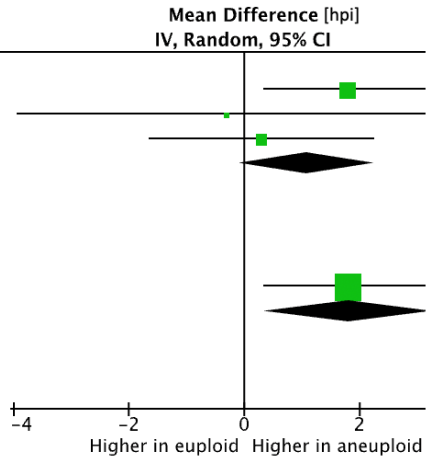


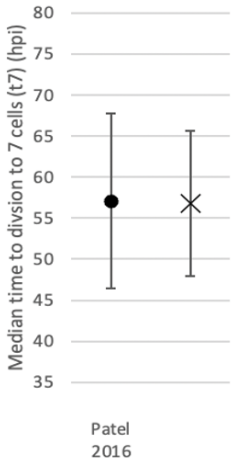


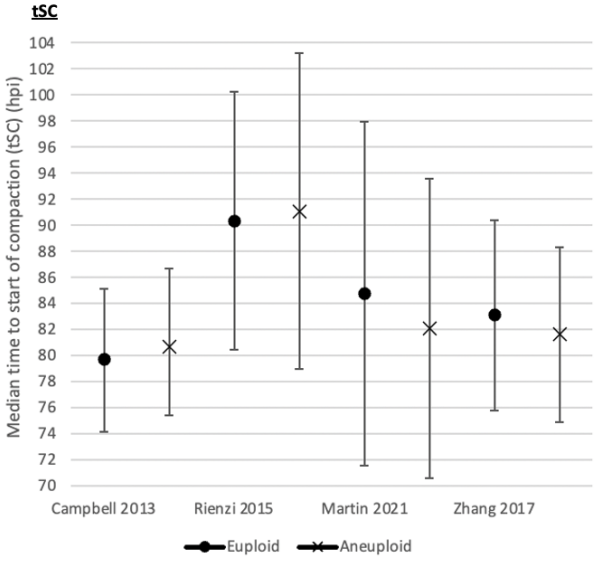
t7

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Mumusoglu 2016	55.3	7.4	242	53.5	7.5	173	56.6%	1.80 [0.34, 3.26]	2016
Patel 2016	56.8	8.9	126	57.1	10.7	41	10.0%	-0.30 [-3.93, 3.33]	2016
Minasi 2016	58.1	16	634	57.8	13.1	297	33.4%	0.30 [-1.64, 2.24]	2016
Subtotal (95% CI)			**1002**			**511**	**100.0%**	**1.09 [-0.07, 2.25]**	
Heterogeneity: Tau² = 0.07; Chi² = 2.11, df = 2 (P = 0.35); I² = 5%									
Test for overall effect: Z = 1.84 (P = 0.07)									
Excluded studies using FISH and D3/blastomere biopsy									
Mumusoglu 2016	55.3	7.4	242	53.5	7.5	173	100.0%	1.80 [0.34, 3.26]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**1.80 [0.34, 3.26]**	
Heterogeneity: Not applicable									
Test for overall effect: Z = 2.42 (P = 0.02)									

Test for subgroup differences: Chi² = 0.56, df = 1 (P = 0.45), I² = 0%







tM

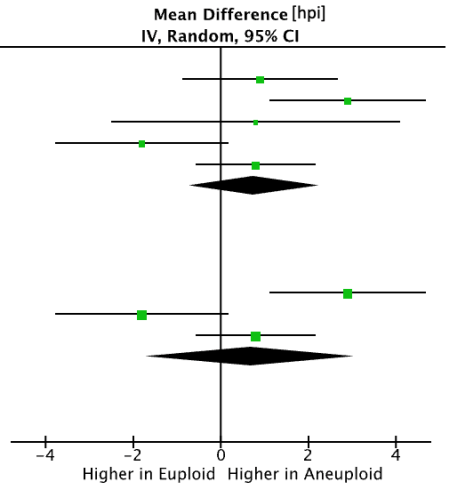
Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Minasi 2016	95.3	14.7	634	94.4	11.8	297	21.5%	0.90 [−0.86, 2.66]	2016
Mumusoglu 2016	94.3	9.2	242	91.4	8.9	173	21.5%	2.90 [1.14, 4.66]	2016
Patel 2016	95.9	10.5	126	95.1	8.9	41	12.2%	0.80 [−2.48, 4.08]	2016
Zhang 2017	86.8	7.6	106	88.6	8.3	150	20.0%	−1.80 [−3.76, 0.16]	2017
Desai 2018	89.6	9.1	394	88.8	9.1	319	24.7%	0.80 [−0.54, 2.14]	2018
Subtotal (95% CI)			**1502**			**980**	**100.0%**	**0.75 [−0.72, 2.23]**	

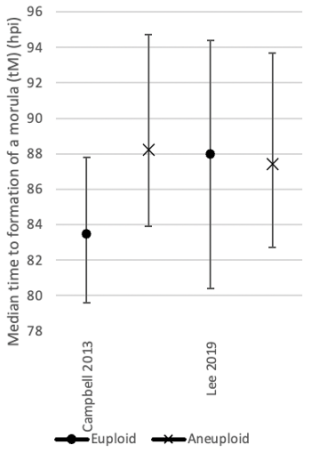
Heterogeneity: $Tau^2 = 1.83$; $Chi^2 = 12.20$, df = 4 (P = 0.02); $I^2 = 67\%$
Test for overall effect: Z = 1.00 (P = 0.32)

Excluded studies using FISH and D3 biopsy									
Mumusoglu 2016	94.3	9.2	242	91.4	8.9	173	32.9%	2.90 [1.14, 4.66]	2016
Zhang 2017	86.8	7.6	106	88.6	8.3	150	31.5%	−1.80 [−3.76, 0.16]	2017
Desai 2018	89.6	9.1	394	88.8	9.1	319	35.6%	0.80 [−0.54, 2.14]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.67 [−1.72, 3.06]**	

Heterogeneity: $Tau^2 = 3.71$; $Chi^2 = 12.19$, df = 2 (P = 0.002); $I^2 = 84\%$
Test for overall effect: Z = 0.55 (P = 0.58)

Test for subgroup differences: $Chi^2 = 0.00$, df = 1 (P = 0.95), $I^2 = 0\%$





tSB

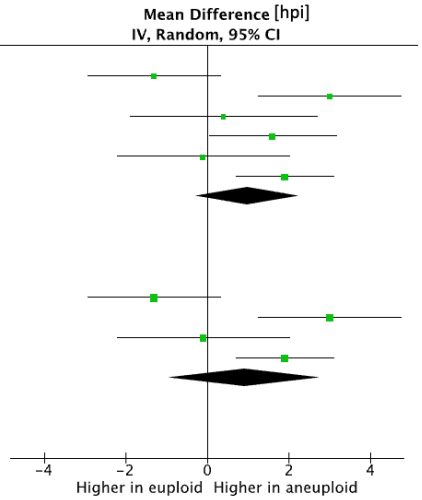
Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Yang 2014	96.1	6.8	135	97.4	6.5	121	17.4%	−1.30 [−2.93, 0.33]	2014
Mumusoglu 2016	102.4	9.3	242	99.4	8.7	173	16.7%	3.00 [1.25, 4.75]	2016
Patel 2016	106.5	7.7	126	106.1	6.1	41	13.5%	0.40 [−1.90, 2.70]	2016
Minasi 2016	105	12.8	634	103.4	10.5	297	17.8%	1.60 [0.04, 3.16]	2016
Zhang 2017	97.7	8.4	106	97.8	8.6	150	14.5%	−0.10 [−2.21, 2.01]	2017
Desai 2018	100.6	7.6	319	98.7	8.6	394	20.1%	1.90 [0.71, 3.09]	2018
Subtotal (95% CI)			**1562**			**1176**	**100.0%**	**0.98 [−0.28, 2.24]**	

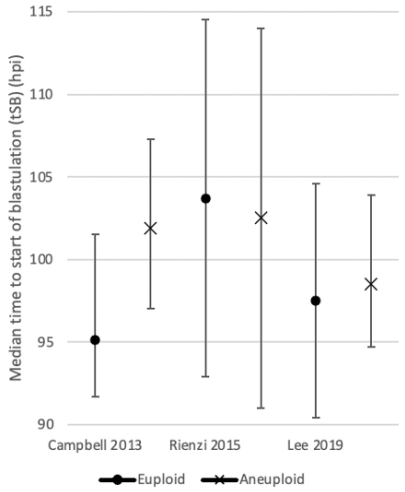
Heterogeneity: $Tau^2 = 1.69$; $Chi^2 = 16.59$, df = 5 (P = 0.005); $I^2 = 70\%$
Test for overall effect: Z = 1.53 (P = 0.13)

Excluded studies using FISH and D3 biopsy									
Yang 2014	96.1	6.8	135	97.4	6.5	121	25.3%	−1.30 [−2.93, 0.33]	2014
Mumusoglu 2016	102.4	9.3	242	99.4	8.7	173	24.6%	3.00 [1.25, 4.75]	2016
Zhang 2017	97.7	8.4	106	97.8	8.6	150	22.4%	−0.10 [−2.21, 2.01]	2017
Desai 2018	100.6	7.6	319	98.7	8.6	394	27.8%	1.90 [0.71, 3.09]	2018
Subtotal (95% CI)			**802**			**838**	**100.0%**	**0.91 [−0.96, 2.79]**	

Heterogeneity: $Tau^2 = 2.92$; $Chi^2 = 15.85$, df = 3 (P = 0.001); $I^2 = 81\%$
Test for overall effect: Z = 0.96 (P = 0.34)

Test for subgroup differences: $Chi^2 = 0.00$, df = 1 (P = 0.95), $I^2 = 0\%$

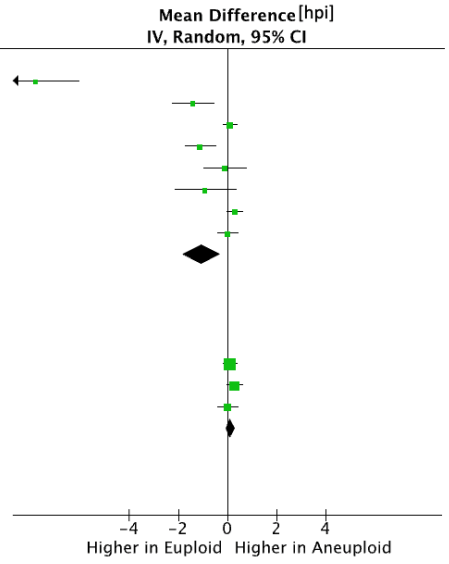


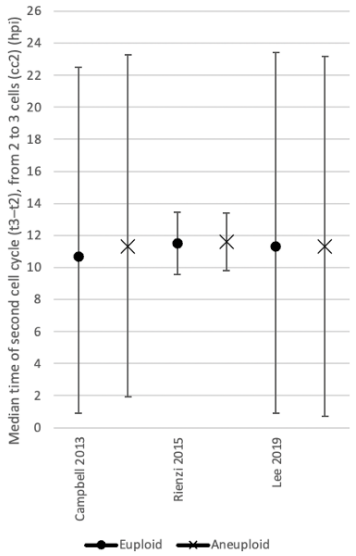


cc2

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Chavez 2012	4	5.2	37	11.8	0.71	8	8.2%	−7.80 [−9.55, −6.05]	2012
Chawla 2014	9.1	4.9	263	10.5	4.2	197	12.5%	−1.40 [−2.23, −0.57]	2014
Yang 2014	11.3	1.1	135	11.2	1.2	121	14.5%	0.10 [−0.18, 0.38]	2014
Basile 2014	10.4	4.4	361	11.5	2.7	143	13.4%	−1.10 [−1.73, −0.47]	2014
Minasi 2016	10.4	6.4	634	10.5	6.1	297	12.4%	−0.10 [−0.95, 0.75]	2016
Patel 2016	10.2	3.8	126	11.1	3.4	41	10.6%	−0.90 [−2.13, 0.33]	2016
Zhang 2017	11	1	106	10.7	1.7	150	14.4%	0.30 [−0.03, 0.63]	2017
Desai 2018	10.8	3.03	394	10.8	2.7	319	14.1%	0.00 [−0.42, 0.42]	2018
Subtotal (95% CI)			**2056**			**1276**	**100.0%**	**−1.01 [−1.76, −0.26]**	
Heterogeneity: $Tau^2 = 0.99$; $Chi^2 = 102.86$, df = 7 ($P < 0.00001$); $I^2 = 93\%$									
Test for overall effect: $Z = 2.64$ ($P = 0.008$)									
Excluded studies using FISH and D3/ blastomere biopsy									
Yang 2014	11.3	1.1	135	11.2	1.2	121	45.9%	0.10 [−0.18, 0.38]	2014
Zhang 2017	11	1	106	10.7	1.7	150	33.4%	0.30 [−0.03, 0.63]	2017
Desai 2018	10.8	3.03	394	10.8	2.7	319	20.7%	0.00 [−0.42, 0.42]	2018
Subtotal (95% CI)			**635**			**590**	**100.0%**	**0.15 [−0.05, 0.34]**	
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 1.39$, df = 2 ($P = 0.50$); $I^2 = 0\%$									
Test for overall effect: $Z = 1.49$ ($P = 0.14$)									

Test for subgroup differences: $Chi^2 = 8.58$, df = 1 ($P = 0.003$), $I^2 = 88.3\%$

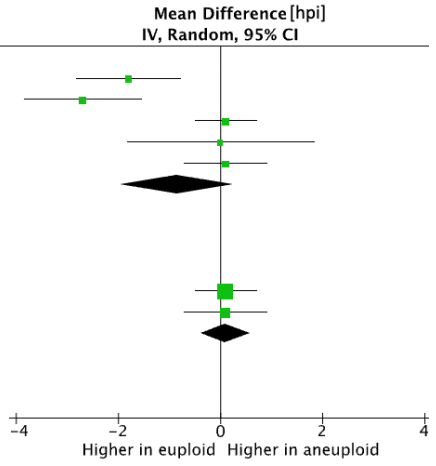


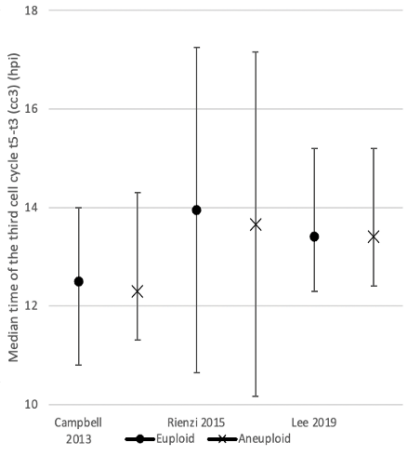


cc3

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Basile 2014	12.2	6.3	361	14	4.8	143	20.5%	−1.80 [−2.82, −0.78]	2014
Chawla 2014	10.8	7	263	13.5	5.6	197	19.6%	−2.70 [−3.85, −1.55]	2014
Mumusoglu 2016	10.4	3.2	242	10.3	3	173	23.1%	0.10 [−0.50, 0.70]	2016
Patel 2016	13.3	6.5	126	13.3	4.7	41	14.9%	0.00 [−1.83, 1.83]	2016
Zhang 2017	13.2	3.4	106	13.1	3.1	150	21.9%	0.10 [−0.72, 0.92]	2017
Subtotal (95% CI)			**1098**			**704**	**100.0%**	**−0.85 [−1.97, 0.26]**	
Heterogeneity: $Tau^2 = 1.31$; $Chi^2 = 26.40$, df = 4 ($P < 0.0001$); $I^2 = 85\%$									
Test for overall effect: $Z = 1.50$ ($P = 0.13$)									
Excluded studies using FISH and D3/ blastomere biopsy									
Mumusoglu 2016	10.4	3.2	242	10.3	3	173	64.7%	0.10 [−0.50, 0.70]	2016
Zhang 2017	13.2	3.4	106	13.1	3.1	150	35.3%	0.10 [−0.72, 0.92]	2017
Subtotal (95% CI)			**348**			**323**	**100.0%**	**0.10 [−0.38, 0.58]**	
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.00$, df = 1 ($P = 1.00$); $I^2 = 0\%$									
Test for overall effect: $Z = 0.40$ ($P = 0.69$)									

Test for subgroup differences: $Chi^2 = 2.36$, df = 1 ($P = 0.12$), $I^2 = 57.6\%$

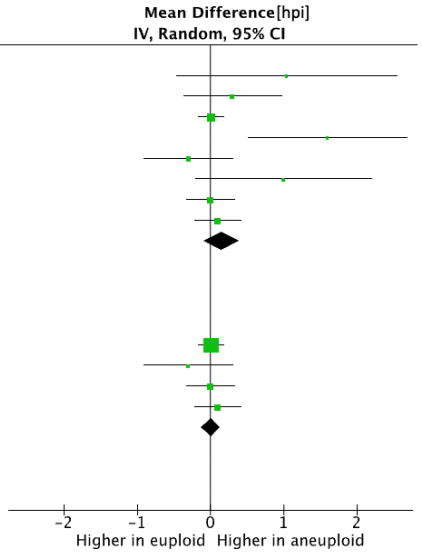


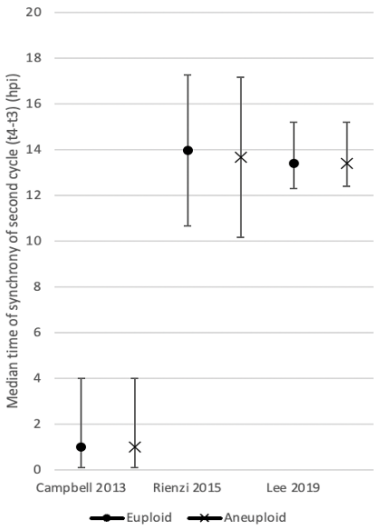


S2

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Chavez 2012	2	4.3	37	0.96	0.84	8	2.5%	1.04 [−0.46, 2.54]	2012
Chawla 2014	2.1	4	263	1.8	3.3	197	9.5%	0.30 [−0.37, 0.97]	2014
Yang 2014	0.78	0.71	135	0.77	0.69	121	27.6%	0.01 [−0.16, 0.18]	2014
Minasi 2016	4.2	7.7	634	2.6	7.9	297	4.5%	1.60 [0.52, 2.68]	2016
Mumusoglu 2016	1.8	3.1	242	2.1	3.1	173	10.9%	−0.30 [−0.90, 0.30]	2016
Patel 2016	2.3	4	126	1.3	3.2	41	3.7%	1.00 [−0.20, 2.20]	2016
Zhang 2017	1.1	1.2	106	1.1	1.5	150	20.2%	0.00 [−0.33, 0.33]	2017
Desai 2018	1.5	1	394	1.4	2.7	319	21.1%	0.10 [−0.21, 0.41]	2018
Subtotal (95% CI)			**1937**			**1306**	**100.0%**	**0.15 [−0.09, 0.40]**	
Heterogeneity: Tau² = 0.05; Chi² = 14.16, df = 7 (P = 0.05); I² = 51%									
Test for overall effect: Z = 1.23 (P = 0.22)									
Excluded studies using FISH and D3/ blastomere biopsy									
Yang 2014	0.78	0.71	135	0.77	0.69	121	60.6%	0.01 [−0.16, 0.18]	2014
Mumusoglu 2016	1.8	3.1	242	2.1	3.1	173	4.9%	−0.30 [−0.90, 0.30]	2016
Zhang 2017	1.1	1.2	106	1.1	1.5	150	16.3%	0.00 [−0.33, 0.33]	2017
Desai 2018	1.5	1	394	1.4	2.7	319	18.3%	0.10 [−0.21, 0.41]	2018
Subtotal (95% CI)			**877**			**763**	**100.0%**	**0.01 [−0.12, 0.14]**	
Heterogeneity: Tau² = 0.00; Chi² = 1.33, df = 3 (P = 0.72); I² = 0%									
Test for overall effect: Z = 0.14 (P = 0.89)									

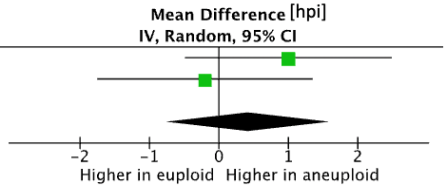
Test for subgroup differences: Chi² = 1.02, df = 1 (P = 0.31), I² = 1.7%

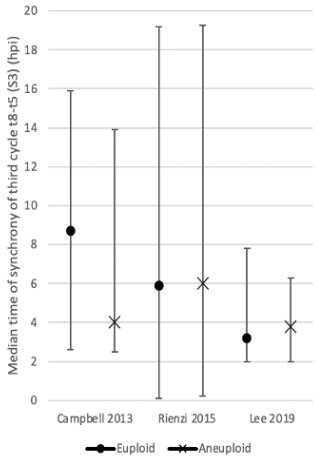




S3

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
Mumusoglu 2016	9.7	8.1	242	8.7	7.2	173	51.7%	1.00 [−0.48, 2.48]	2016
Zhang 2017	6.5	5.9	106	6.7	6.6	150	48.3%	−0.20 [−1.74, 1.34]	2017
Total (95% CI)			**348**			**323**	**100.0%**	**0.42 [−0.76, 1.60]**	
Heterogeneity: Tau² = 0.13; Chi² = 1.21, df = 1 (P = 0.27); I² = 17%									
Test for overall effect: Z = 0.70 (P = 0.48)									





t5-t2

Study or Subgroup	Aneuploid Mean	SD	Total	Euploid Mean	SD	Total	Weight	Mean Difference IV, Random, 95% CI	Year
All studies									
Chawla 2014	20.1	8.6	263	23.9	7.4	197	16.3%	-3.80 [-5.27, -2.33]	2014
Basile 2014	22.6	7.2	361	25.5	5.5	143	17.3%	-2.90 [-4.07, -1.73]	2014
Patel 2016	23.7	7.2	126	24.4	5.7	41	13.9%	-0.70 [-2.85, 1.45]	2016
Mumusoglu 2016	22.5	6.3	242	22.2	6	173	17.2%	0.30 [-0.90, 1.50]	2016
Zhang 2017	24.2	3.8	106	23.8	4.1	150	17.9%	0.40 [-0.58, 1.38]	2017
Desai 2018	23.7	5.5	394	23.3	9.5	319	17.3%	0.40 [-0.78, 1.58]	2018
Subtotal (95% CI)			**1492**			**1023**	**100.0%**	**-1.03 [-2.53, 0.47]**	

Heterogeneity: $Tau^2 = 3.04$; $Chi^2 = 41.58$, df = 5 ($P < 0.00001$); $I^2 = 88\%$

Test for overall effect: Z = 1.34 (P = 0.18)

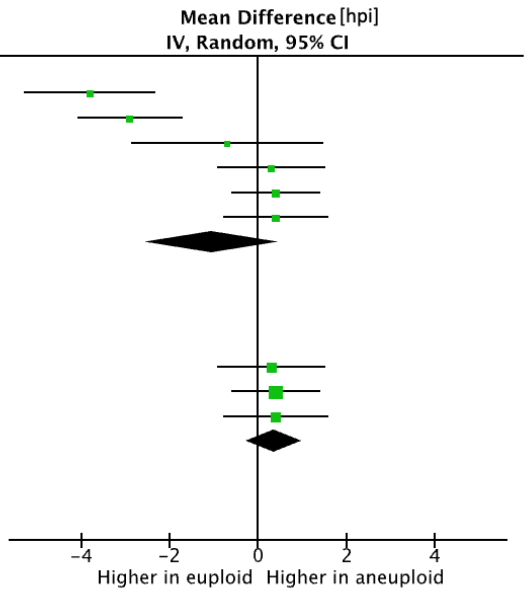
Excluded studies using FISH and D3/ blastomere biopsy

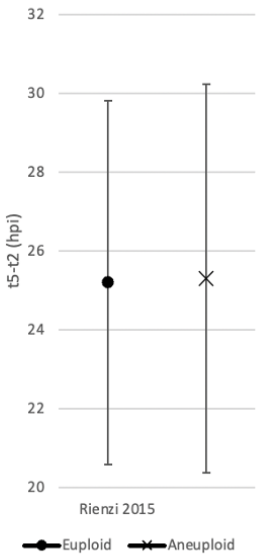
Mumusoglu 2016	22.5	6.3	242	22.2	6	173	28.3%	0.30 [-0.90, 1.50]	2016
Zhang 2017	24.2	3.8	106	23.8	4.1	150	42.4%	0.40 [-0.58, 1.38]	2017
Desai 2018	23.7	5.5	394	23.3	9.5	319	29.3%	0.40 [-0.78, 1.58]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.37 [-0.26, 1.01]**	

Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.02$, df = 2 ($P = 0.99$); $I^2 = 0\%$

Test for overall effect: Z = 1.15 (P = 0.25)

Test for subgroup differences: $Chi^2 = 2.83$, df = 1 (P = 0.09), $I^2 = 64.6\%$





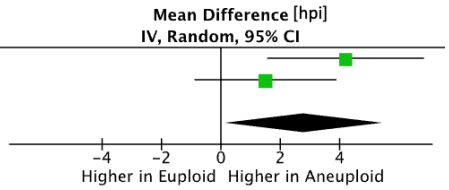
tHB

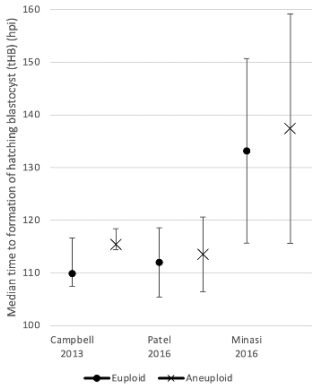
Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Random, 95% CI	Year
Minasi 2016	137.4	21.8	634	133.2	17.5	297	47.8%	4.20 [1.58, 6.82]	2016
Patel 2016	113.5	7.1	126	112	6.6	41	52.2%	1.50 [-0.87, 3.87]	2016
Total (95% CI)			**760**			**338**	**100.0%**	**2.79 [0.15, 5.43]**	

Heterogeneity: $Tau^2 = 2.02$; $Chi^2 = 2.25$, df = 1 ($P = 0.13$); $I^2 = 56\%$

Test for overall effect: Z = 2.07 (P = 0.04)

Excluded studies using FISH and D3/ blastomere biopsy



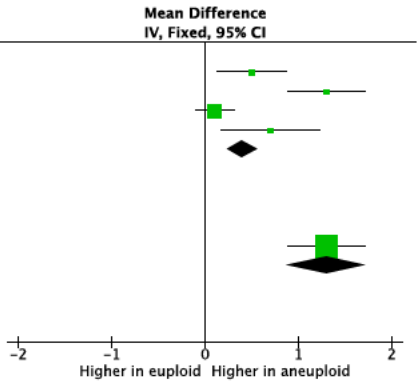


Supplementary Figure 3: Sensitivity Analysis for Morphokinetics using a fixed effects model

tPB2

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
1.1.1 All studies									
Chawla 2014	4.5	1.8	263	4	2.2	197	18.0%	0.50 [0.12, 0.88]	2014
Mumusoglu 2016	4.1	1.9	242	2.8	2.3	173	14.6%	1.30 [0.88, 1.72]	2016
Minasi 2016	4	1.9	634	3.9	1.3	297	58.4%	0.10 [−0.11, 0.31]	2016
Patel 2016	4	1.7	41	3.3	0.7	126	8.9%	0.70 [0.17, 1.23]	2016
Subtotal (95% CI)			**1180**			**793**	**100.0%**	**0.40 [0.24, 0.56]**	
Heterogeneity: $Chi^2 = 27.19$, df = 3 ($P < 0.00001$); $I^2 = 89\%$									
Test for overall effect: $Z = 4.92$ ($P < 0.00001$)									
1.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	4.1	1.9	242	2.8	2.3	173	100.0%	1.30 [0.88, 1.72]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**1.30 [0.88, 1.72]**	
Heterogeneity: Not applicable									
Test for overall effect: $Z = 6.09$ ($P < 0.00001$)									

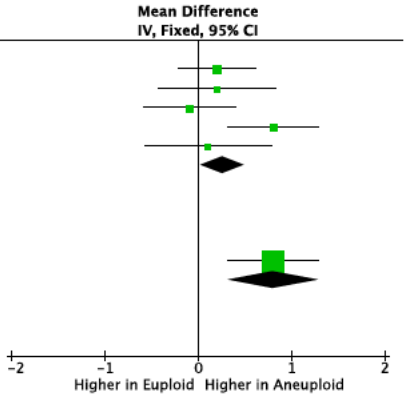
Test for subgroup differences: $Chi^2 = 15.49$, df = 1 ($P < 0.0001$), $I^2 = 93.5\%$



tPN

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
2.1.1 All studies									
Basile 2014	9	2.9	361	8.8	1.8	143	30.2%	0.20 [−0.22, 0.62]	2014
Chawla 2014	10.8	3.4	263	10.6	3.4	197	13.5%	0.20 [−0.43, 0.83]	2014
Minasi 2016	10.1	3.8	634	10.2	3.5	297	21.7%	−0.10 [−0.60, 0.40]	2016
Mumusoglu 2016	9.4	3	242	8.6	2	173	23.0%	0.80 [0.32, 1.28]	2016
Patel 2016	7.7	2.5	126	7.6	1.7	41	11.6%	0.10 [−0.58, 0.78]	2016
Subtotal (95% CI)			**1626**			**851**	**100.0%**	**0.26 [0.03, 0.49]**	
Heterogeneity: $Chi^2 = 7.19$, df = 4 ($P = 0.13$); $I^2 = 44\%$									
Test for overall effect: $Z = 2.22$ ($P = 0.03$)									
2.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	9.4	3	242	8.6	2	173	100.0%	0.80 [0.32, 1.28]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**0.80 [0.32, 1.28]**	
Heterogeneity: Not applicable									
Test for overall effect: $Z = 3.26$ ($P = 0.001$)									

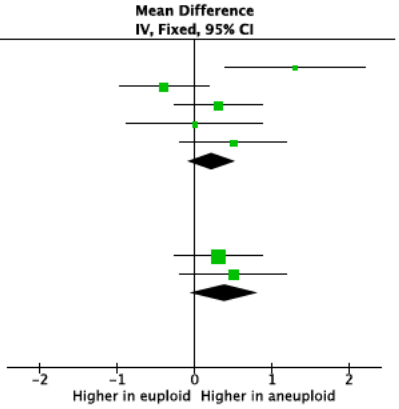
Test for subgroup differences: $Chi^2 = 3.91$, df = 1 ($P = 0.05$), $I^2 = 74.4\%$



tPNf

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
3.1.1 All studies									
Chawla 2014	25.8	5.6	263	24.5	4.3	197	11.5%	1.30 [0.40, 2.20]	2014
Minasi 2016	24.4	3.5	634	24.8	4.5	297	27.9%	−0.40 [−0.98, 0.18]	2016
Mumusoglu 2016	24.8	3.1	242	24.5	2.8	173	28.7%	0.30 [−0.27, 0.87]	2016
Patel 2016	7.7	2.5	126	7.7	2.5	41	12.1%	0.00 [−0.88, 0.88]	2016
Zhang 2017	22.5	3	106	22	2.4	150	19.8%	0.50 [−0.19, 1.19]	2017
Subtotal (95% CI)			**1371**			**858**	**100.0%**	**0.22 [−0.08, 0.53]**	
Heterogeneity: $Chi^2 = 10.82$, df = 4 ($P = 0.03$); $I^2 = 63\%$									
Test for overall effect: $Z = 1.42$ ($P = 0.15$)									
3.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	24.8	3.1	242	24.5	2.8	173	59.2%	0.30 [−0.27, 0.87]	2016
Zhang 2017	22.5	3	106	22	2.4	150	40.8%	0.50 [−0.19, 1.19]	2017
Subtotal (95% CI)			**348**			**323**	**100.0%**	**0.38 [−0.06, 0.82]**	
Heterogeneity: $Chi^2 = 0.19$, df = 1 ($P = 0.66$); $I^2 = 0\%$									
Test for overall effect: $Z = 1.70$ ($P = 0.09$)									

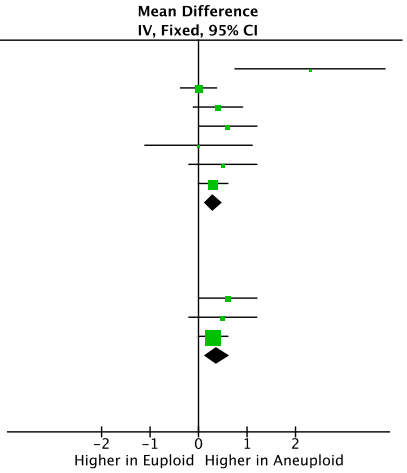
Test for subgroup differences: $Chi^2 = 0.34$, df = 1 ($P = 0.56$), $I^2 = 0\%$



t2

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
4.1.1 All studies									
Chawla 2014	30.6	9.7	263	28.3	7.2	197	1.5%	2.30 [0.76, 3.84]	2014
Basile 2014	26.8	3.4	361	26.8	0.7	143	25.9%	0.00 [-0.37, 0.37]	2014
Minasi 2016	28.4	0.9	634	28	4.4	297	13.8%	0.40 [-0.11, 0.91]	2016
Mumusoglu 2016	27.3	3.3	242	26.7	2.9	173	9.8%	0.60 [0.00, 1.20]	2016
Patel 2016	27.6	3.5	126	27.6	3	41	2.9%	0.00 [-1.10, 1.10]	2016
Zhang 2017	25	3	106	24.5	2.5	150	7.2%	0.50 [-0.20, 1.20]	2017
Desai 2018	26.2	0.3	394	25.9	2.73	319	38.9%	0.30 [-0.00, 0.60]	2018
Subtotal (95% CI)			**2126**			**1320**	**100.0%**	**0.30 [0.11, 0.49]**	
Heterogeneity: Chi² = 10.69, df = 6 (P = 0.10); I² = 44%									
Test for overall effect: Z = 3.14 (P = 0.002)									
4.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	27.3	3.3	242	26.7	2.9	173	17.5%	0.60 [0.00, 1.20]	2016
Zhang 2017	25	3	106	24.5	2.5	150	13.0%	0.50 [-0.20, 1.20]	2017
Desai 2018	26.2	0.3	394	25.9	2.73	319	69.5%	0.30 [-0.00, 0.60]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.38 [0.13, 0.63]**	
Heterogeneity: Chi² = 0.90, df = 2 (P = 0.64); I² = 0%									
Test for overall effect: Z = 2.96 (P = 0.003)									

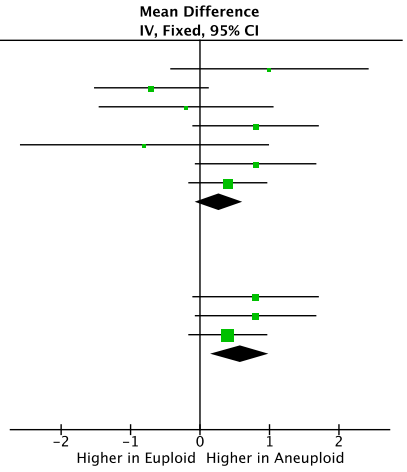
Test for subgroup differences: Chi² = 0.24, df = 1 (P = 0.63), I² = 0%



t3

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
5.1.1 All studies									
Chawla 2014	39.7	8.5	263	38.7	7	197	5.7%	1.00 [-0.42, 2.42]	2014
Basile 2014	37.2	4.8	361	37.9	3.96	143	17.2%	-0.70 [-1.52, 0.12]	2014
Minasi 2016	37.2	8.99	297	37.4	9.23	634	7.3%	-0.20 [-1.45, 1.05]	2016
Mumusoglu 2016	37.6	4.5	242	36.8	4.7	173	14.1%	0.80 [-0.10, 1.70]	2016
Patel 2016	37.9	5.2	41	38.7	4.6	126	3.6%	-0.80 [-2.58, 0.98]	2016
Zhang 2017	36	3.6	106	35.2	3.3	150	15.3%	0.80 [-0.07, 1.67]	2017
Desai 2018	37.1	4	394	36.7	3.6	319	36.7%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**1704**			**1742**	**100.0%**	**0.28 [-0.06, 0.61]**	
Heterogeneity: Chi² = 11.35, df = 6 (P = 0.08); I² = 47%									
Test for overall effect: Z = 1.59 (P = 0.11)									
5.1.2 Exclude FISH and D3/blastomere biopsy									
Mumusoglu 2016	37.6	4.5	242	36.8	4.7	173	21.3%	0.80 [-0.10, 1.70]	2016
Zhang 2017	36	3.6	106	35.2	3.3	150	23.1%	0.80 [-0.07, 1.67]	2017
Desai 2018	37.1	4	394	36.7	3.6	319	55.5%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.58 [0.16, 0.99]**	
Heterogeneity: Chi² = 0.88, df = 2 (P = 0.65); I² = 0%									
Test for overall effect: Z = 2.72 (P = 0.007)									

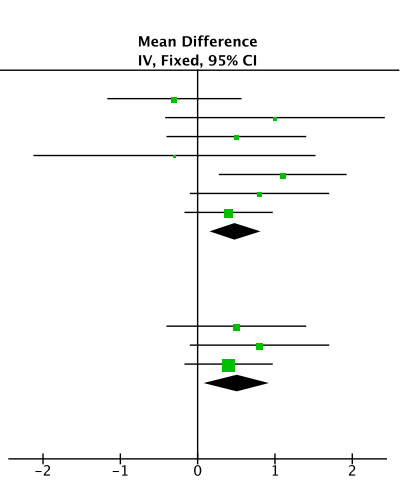
Test for subgroup differences: Chi² = 1.22, df = 1 (P = 0.27), I² = 18.1%



t4

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
6.1.1 All studies									
Basile 2014	39.6	4.84	361	39.9	4.27	143	14.5%	-0.30 [-1.16, 0.56]	2014
Chawla 2014	41.5	8.2	263	40.5	7.2	197	5.4%	1.00 [-0.41, 2.41]	2014
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	13.4%	0.50 [-0.39, 1.39]	2016
Patel 2016	40.1	5	126	40.4	5.2	41	3.3%	-0.30 [-2.12, 1.52]	2016
Minasi 2016	41.1	6.4	634	40	5.7	297	16.0%	1.10 [0.28, 1.92]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	13.4%	0.80 [-0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	34.0%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**2126**			**1320**	**100.0%**	**0.49 [0.16, 0.81]**	
Heterogeneity: Chi² = 7.17, df = 6 (P = 0.31); I² = 16%									
Test for overall effect: Z = 2.92 (P = 0.004)									
6.1.2 Excluded FISH and D3/ blastomere biopsy									
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	22.0%	0.50 [-0.39, 1.39]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	22.1%	0.80 [-0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	55.9%	0.40 [-0.16, 0.96]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.51 [0.09, 0.93]**	
Heterogeneity: Chi² = 0.55, df = 2 (P = 0.76); I² = 0%									
Test for overall effect: Z = 2.38 (P = 0.02)									

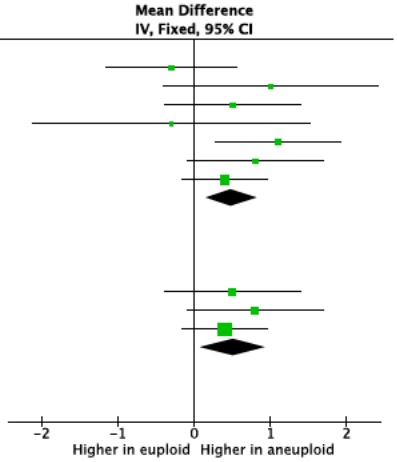
Test for subgroup differences: Chi² = 0.01, df = 1 (P = 0.93), I² = 0%



t5

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
6.1.1 All studies									
Basile 2014	39.6	4.84	361	39.9	4.27	143	14.5%	−0.30 [−1.16, 0.56]	2014
Chawla 2014	41.5	8.2	263	40.5	7.2	197	5.4%	1.00 [−0.41, 2.41]	2014
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	13.4%	0.50 [−0.39, 1.39]	2016
Patel 2016	40.1	5	126	40.4	5.2	41	3.3%	−0.30 [−2.12, 1.52]	2016
Minasi 2016	41.1	6.4	634	40	5.7	297	16.0%	1.10 [0.28, 1.92]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	13.4%	0.80 [−0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	34.0%	0.40 [−0.16, 0.96]	2018
Subtotal (95% CI)			**2126**			**1320**	**100.0%**	**0.49 [0.16, 0.81]**	
Heterogeneity: $Chi^2 = 7.17$, df = 6 (P = 0.31); $I^2 = 16\%$									
Test for overall effect: Z = 2.92 (P = 0.004)									
6.1.2 Excluded FISH and D3/ blastomere biopsy									
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	22.0%	0.50 [−0.39, 1.39]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	22.1%	0.80 [−0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	55.9%	0.40 [−0.16, 0.96]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.51 [0.09, 0.93]**	
Heterogeneity: $Chi^2 = 0.55$, df = 2 (P = 0.76); $I^2 = 0\%$									
Test for overall effect: Z = 2.38 (P = 0.02)									

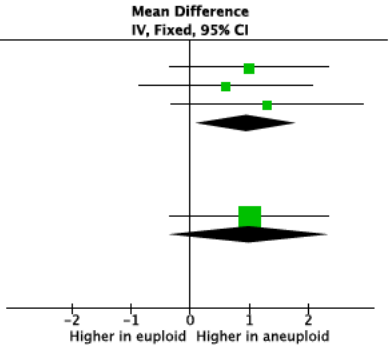
Test for subgroup differences: $Chi^2 = 0.01$, df = 1 (P = 0.93), $I^2 = 0\%$



t6

Study or Subgroup	Aneuploidy Mean [hpi]	Aneuploidy SD [hpi]	Aneuploidy Total	Euploidy Mean [hpi]	Euploidy SD [hpi]	Euploidy Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
8.1.1 All studies									
Mumusoglu 2016	52.4	7.2	242	51.4	6.6	173	39.8%	1.00 [−0.34, 2.34]	2016
Patel 2016	54	7.9	126	53.4	6.1	634	33.5%	0.60 [−0.86, 2.06]	2016
Minasi 2016	55.2	13.5	634	53.9	11	297	26.7%	1.30 [−0.33, 2.93]	2016
Subtotal (95% CI)			**1002**			**1104**	**100.0%**	**0.95 [0.10, 1.79]**	
Heterogeneity: $Chi^2 = 0.40$, df = 2 (P = 0.82); $I^2 = 0\%$									
Test for overall effect: Z = 2.20 (P = 0.03)									
8.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	52.4	7.2	242	51.4	6.6	173	100.0%	1.00 [−0.34, 2.34]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**1.00 [−0.34, 2.34]**	
Heterogeneity: Not applicable									
Test for overall effect: Z = 1.46 (P = 0.14)									

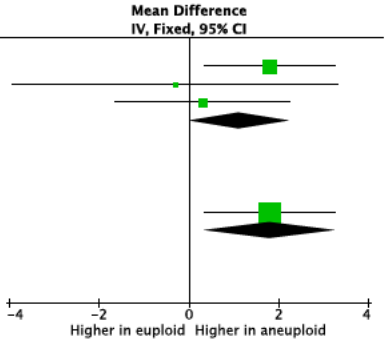
Test for subgroup differences: $Chi^2 = 0.00$, df = 1 (P = 0.95), $I^2 = 0\%$



t7

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
9.1.1 All									
Mumusoglu 2016	55.3	7.4	242	53.5	7.5	173	58.0%	1.80 [0.34, 3.26]	2016
Patel 2016	56.8	8.9	126	57.1	10.7	41	9.4%	−0.30 [−3.93, 3.33]	2016
Minasi 2016	58.1	16	634	57.8	13.1	297	32.6%	0.30 [−1.64, 2.24]	2016
Subtotal (95% CI)			**1002**			**511**	**100.0%**	**1.11 [0.01, 2.22]**	
Heterogeneity: $Chi^2 = 2.11$, df = 2 (P = 0.35); $I^2 = 5\%$									
Test for overall effect: Z = 1.97 (P = 0.05)									
9.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	55.3	7.4	242	53.5	7.5	173	100.0%	1.80 [0.34, 3.26]	2016
Subtotal (95% CI)			**242**			**173**	**100.0%**	**1.80 [0.34, 3.26]**	
Heterogeneity: Not applicable									
Test for overall effect: Z = 2.42 (P = 0.02)									

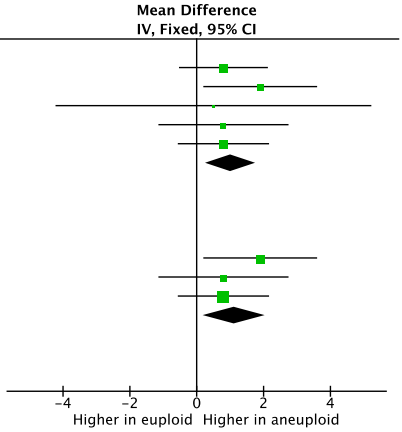
Test for subgroup differences: $Chi^2 = 0.54$, df = 1 (P = 0.46), $I^2 = 0\%$



t8

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
10.1.1 All studies									
Minasi 2016	62	12.8	634	61.2	7.47	297	32.3%	0.80 [−0.51, 2.11]	2016
Mumusoglu 2016	58.8	9.3	242	56.9	8.1	173	19.6%	1.90 [0.22, 3.58]	2016
Patel 2016	61.7	13.5	126	61.2	13.3	41	2.5%	0.50 [−4.20, 5.20]	2016
Zhang 2017	55.7	7.7	106	54.9	7.8	150	14.9%	0.80 [−1.13, 2.73]	2017
Desai 2018	59.3	8.6	394	58.5	9.5	319	30.6%	0.80 [−0.54, 2.14]	2018
Subtotal (95% CI)			**1502**			**980**	**100.0%**	**1.01 [0.26, 1.75]**	
Heterogeneity: Chi² = 1.36, df = 4 (P = 0.85); I² = 0%									
Test for overall effect: Z = 2.65 (P = 0.008)									
10.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	58.8	9.3	242	56.9	8.1	173	30.0%	1.90 [0.22, 3.58]	2016
Zhang 2017	55.7	7.7	106	54.9	7.8	150	22.9%	0.80 [−1.13, 2.73]	2017
Desai 2018	59.3	8.6	394	58.5	9.5	319	47.0%	0.80 [−0.54, 2.14]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**1.13 [0.21, 2.05]**	
Heterogeneity: Chi² = 1.15, df = 2 (P = 0.56); I² = 0%									
Test for overall effect: Z = 2.40 (P = 0.02)									

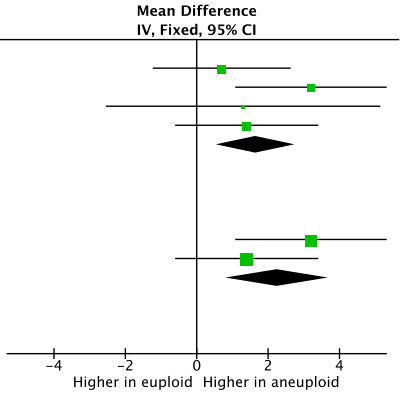
Test for subgroup differences: Chi² = 0.04, df = 1 (P = 0.84), I² = 0%



t9

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
11.1.1 All studies									
Minasi 2016	76.8	15.4	634	76.1	13.1	297	33.3%	0.70 [−1.21, 2.61]	2016
Mumusoglu 2016	71.6	11.4	242	68.4	10.3	173	27.5%	3.20 [1.10, 5.30]	2016
Patel 2016	73.5	12.4	126	72.2	10.3	41	8.3%	1.30 [−2.52, 5.12]	2016
Zhang 2017	70.4	7.6	106	69	8.5	150	30.9%	1.40 [−0.59, 3.39]	2017
Subtotal (95% CI)			**1108**			**661**	**100.0%**	**1.65 [0.55, 2.76]**	
Heterogeneity: Chi² = 3.13, df = 3 (P = 0.37); I² = 4%									
Test for overall effect: Z = 2.94 (P = 0.003)									
11.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	71.6	11.4	242	68.4	10.3	173	47.2%	3.20 [1.10, 5.30]	2016
Zhang 2017	70.4	7.6	106	69	8.5	150	52.8%	1.40 [−0.59, 3.39]	2017
Subtotal (95% CI)			**348**			**323**	**100.0%**	**2.25 [0.81, 3.69]**	
Heterogeneity: Chi² = 1.49, df = 1 (P = 0.22); I² = 33%									
Test for overall effect: Z = 3.05 (P = 0.002)									

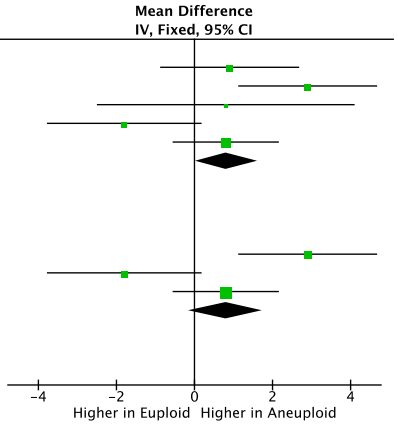
Test for subgroup differences: Chi² = 0.41, df = 1 (P = 0.52), I² = 0%



tM

Study or Subgroup	Aneuploid Mean [hpi]	SD [hpi]	Total	Euploid Mean [hpi]	SD [hpi]	Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
12.1.1 All studies									
Minasi 2016	95.3	14.7	634	94.4	11.8	297	20.7%	0.90 [−0.86, 2.66]	2016
Mumusoglu 2016	94.3	9.2	242	91.4	8.9	173	20.8%	2.90 [1.14, 4.66]	2016
Patel 2016	95.9	10.5	126	95.1	8.9	41	6.0%	0.80 [−2.48, 4.08]	2016
Zhang 2017	86.8	7.6	106	88.6	8.3	150	16.7%	−1.80 [−3.76, 0.16]	2017
Desai 2018	89.6	9.1	394	88.8	9.1	319	35.8%	0.80 [−0.54, 2.14]	2018
Subtotal (95% CI)			**1502**			**980**	**100.0%**	**0.82 [0.02, 1.63]**	
Heterogeneity: Chi² = 12.20, df = 4 (P = 0.02); I² = 67%									
Test for overall effect: Z = 2.01 (P = 0.04)									
12.1.2 Excluding FISH and D3/ blastomere biopsy									
Mumusoglu 2016	94.3	9.2	242	91.4	8.9	173	28.4%	2.90 [1.14, 4.66]	2016
Zhang 2017	86.8	7.6	106	88.6	8.3	150	22.8%	−1.80 [−3.76, 0.16]	2017
Desai 2018	89.6	9.1	394	88.8	9.1	319	48.8%	0.80 [−0.54, 2.14]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.80 [−0.14, 1.74]**	
Heterogeneity: Chi² = 12.19, df = 2 (P = 0.002); I² = 84%									
Test for overall effect: Z = 1.68 (P = 0.09)									

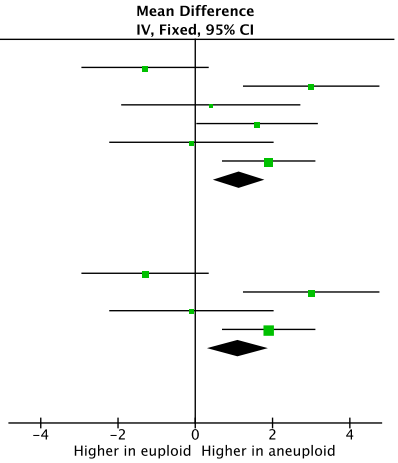
Test for subgroup differences: Chi² = 0.00, df = 1 (P = 0.97), I² = 0%



tSB

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
13.1.1 All studies									
Yang 2014	96.1	6.8	135	97.4	6.5	121	16.8%	−1.30 [−2.93, 0.33]	2014
Mumusoglu 2016	102.4	9.3	242	99.4	8.7	173	14.6%	3.00 [1.25, 4.75]	2016
Patel 2016	106.5	7.7	126	106.1	6.1	41	8.4%	0.40 [−1.90, 2.70]	2016
Minasi 2016	105	12.8	634	103.4	10.5	297	18.5%	1.60 [0.04, 3.16]	2016
Zhang 2017	97.7	8.4	106	97.8	8.6	150	10.0%	−0.10 [−2.21, 2.01]	2017
Desai 2018	100.6	7.6	319	98.7	8.6	394	31.6%	1.90 [0.71, 3.09]	2018
Subtotal (95% CI)			**1562**			**1176**	**100.0%**	**1.14 [0.47, 1.81]**	
Heterogeneity: Chi² = 16.59, df = 5 (P = 0.005); I² = 70%									
Test for overall effect: Z = 3.34 (P = 0.0008)									
13.1.2 Excluding FISH and D3/blastomere biopsy									
Yang 2014	96.1	6.8	135	97.4	6.5	121	23.0%	−1.30 [−2.93, 0.33]	2014
Mumusoglu 2016	102.4	9.3	242	99.4	8.7	173	20.0%	3.00 [1.25, 4.75]	2016
Zhang 2017	97.7	8.4	106	97.8	8.6	150	13.7%	−0.10 [−2.21, 2.01]	2017
Desai 2018	100.6	7.6	319	98.7	8.6	394	43.2%	1.90 [0.71, 3.09]	2018
Subtotal (95% CI)			**802**			**838**	**100.0%**	**1.11 [0.33, 1.89]**	
Heterogeneity: Chi² = 15.85, df = 3 (P = 0.001); I² = 81%									
Test for overall effect: Z = 2.78 (P = 0.005)									

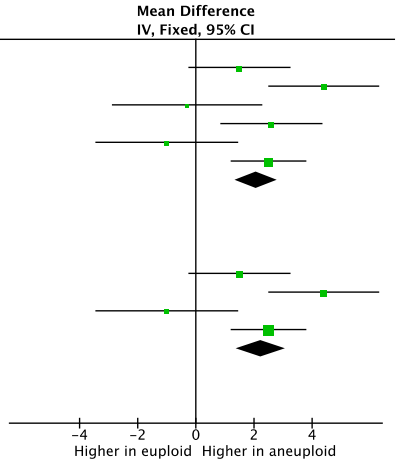
Test for subgroup differences: Chi² = 0.00, df = 1 (P = 0.95), I² = 0%



tB

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
14.1.1 All patients									
Yang 2014	104.3	6.9	135	102.8	7.2	121	17.7%	1.50 [−0.23, 3.23]	2014
Mumusoglu 2016	112.5	10.4	242	108.1	9.1	173	15.0%	4.40 [2.51, 6.29]	2016
Patel 2016	109.7	7.2	126	110	7.3	41	8.1%	−0.30 [−2.86, 2.26]	2016
Minasi 2016	112.8	14.1	634	110.2	11.8	297	17.7%	2.60 [0.87, 4.33]	2016
Zhang 2017	106.5	9.7	106	107.5	9.9	150	9.0%	−1.00 [−3.43, 1.43]	2017
Desai 2018	106.4	8.1	394	103.9	9.1	319	32.5%	2.50 [1.22, 3.78]	2018
Subtotal (95% CI)			**1637**			**1101**	**100.0%**	**2.08 [1.35, 2.81]**	
Heterogeneity: Chi² = 16.47, df = 5 (P = 0.006); I² = 70%									
Test for overall effect: Z = 5.60 (P < 0.00001)									
14.1.2 Excluding FISH and D3/blastomere biopsy									
Yang 2014	104.3	6.9	135	102.8	7.2	121	23.9%	1.50 [−0.23, 3.23]	2014
Mumusoglu 2016	112.5	10.4	242	108.1	9.1	173	20.2%	4.40 [2.51, 6.29]	2016
Zhang 2017	106.5	9.7	106	107.5	9.9	150	12.1%	−1.00 [−3.43, 1.43]	2017
Desai 2018	106.4	8.1	394	103.9	9.1	319	43.8%	2.50 [1.22, 3.78]	2018
Subtotal (95% CI)			**877**			**763**	**100.0%**	**2.22 [1.37, 3.07]**	
Heterogeneity: Chi² = 12.71, df = 3 (P = 0.005); I² = 76%									
Test for overall effect: Z = 5.14 (P < 0.00001)									

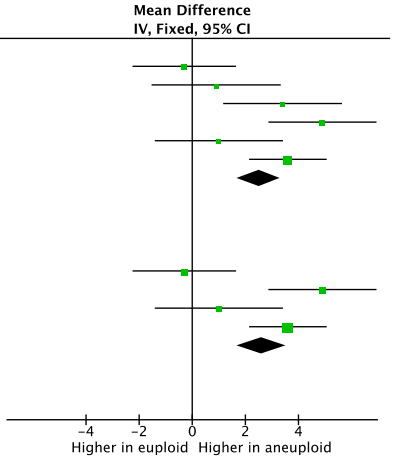
Test for subgroup differences: Chi² = 0.06, df = 1 (P = 0.81), I² = 0%



tEB

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
15.1.1 All patients									
Yang 2014	110.9	8.1	135	111.2	7.6	121	17.4%	−0.30 [−2.22, 1.62]	2014
Patel 2016	111.6	5.1	126	110.7	7.3	41	11.1%	0.90 [−1.51, 3.31]	2016
Minasi 2016	122.1	17.3	634	118.7	15.4	297	13.2%	3.40 [1.19, 5.61]	2016
Mumusoglu 2016	118.7	11.2	242	113.8	9.6	173	15.9%	4.90 [2.89, 6.91]	2016
Zhang 2017	121.5	9.6	106	120.5	9.6	150	11.3%	1.00 [−1.39, 3.39]	2017
Desai 2018	113.9	9.1	394	110.3	10.2	319	31.2%	3.60 [2.16, 5.04]	2018
Subtotal (95% CI)			**1637**			**1101**	**100.0%**	**2.51 [1.71, 3.31]**	
Heterogeneity: Chi² = 19.73, df = 5 (P = 0.001); I² = 75%									
Test for overall effect: Z = 6.14 (P < 0.00001)									
15.1.2 Excluding FISH and D3/blastomere biopsy									
Yang 2014	110.9	8.1	135	111.2	7.6	121	22.9%	−0.30 [−2.22, 1.62]	2014
Mumusoglu 2016	118.7	11.2	242	113.8	9.6	173	21.0%	4.90 [2.89, 6.91]	2016
Zhang 2017	121.5	9.6	106	120.5	9.6	150	14.9%	1.00 [−1.39, 3.39]	2017
Desai 2018	113.9	9.1	394	110.3	10.2	319	41.2%	3.60 [2.16, 5.04]	2018
Subtotal (95% CI)			**877**			**763**	**100.0%**	**2.59 [1.67, 3.51]**	
Heterogeneity: Chi² = 17.35, df = 3 (P = 0.0006); I² = 83%									
Test for overall effect: Z = 5.52 (P < 0.00001)									

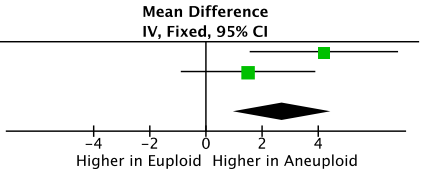
Test for subgroup differences: Chi² = 0.02, df = 1 (P = 0.90), I² = 0%



tHB

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
Minasi 2016	137.4	21.8	634	133.2	17.5	297	45.1%	4.20 [1.58, 6.82]	2016
Patel 2016	113.5	7.1	126	112	6.6	41	54.9%	1.50 [-0.87, 3.87]	2016
Total (95% CI)			**760**			**338**	**100.0%**	**2.72 [0.96, 4.47]**	

Heterogeneity: $Chi^2 = 2.25$, df = 1 (P = 0.13); $I^2 = 56\%$
Test for overall effect: Z = 3.03 (P = 0.002)



cc2

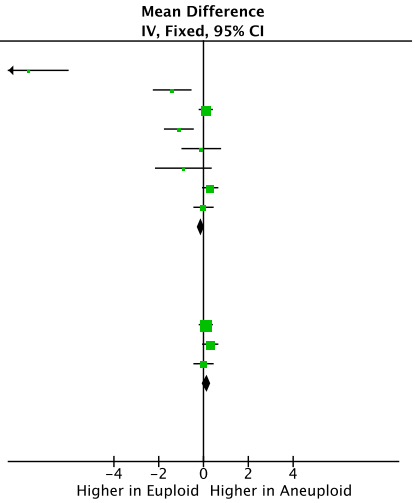
Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
17.1.1 All studies									
Chavez 2012	4	5.2	37	11.8	0.71	8	1.0%	-7.80 [-9.55, -6.05]	2012
Chawla 2014	9.1	4.9	263	10.5	4.2	197	4.3%	-1.40 [-2.23, -0.57]	2014
Yang 2014	11.3	1.1	135	11.2	1.2	121	37.3%	0.10 [-0.18, 0.38]	2014
Basile 2014	10.4	4.4	361	11.5	2.7	143	7.4%	-1.10 [-1.73, -0.47]	2014
Minasi 2016	10.4	6.4	634	10.5	6.1	297	4.1%	-0.10 [-0.95, 0.75]	2016
Patel 2016	10.2	3.8	126	11.1	3.4	41	2.0%	-0.90 [-2.13, 0.33]	2016
Zhang 2017	11	1	106	10.7	1.7	150	27.1%	0.30 [-0.03, 0.63]	2017
Desai 2018	10.8	3.03	394	10.8	2.7	319	16.9%	0.00 [-0.42, 0.42]	2018
Subtotal (95% CI)			**2056**			**1276**	**100.0%**	**-0.12 [-0.29, 0.05]**	

Heterogeneity: $Chi^2 = 102.86$, df = 7 (P < 0.00001); $I^2 = 93\%$
Test for overall effect: Z = 1.38 (P = 0.17)

17.1.2 Excluding FISH and D3/blastomere biopsy									
Yang 2014	11.3	1.1	135	11.2	1.2	121	45.9%	0.10 [-0.18, 0.38]	2014
Zhang 2017	11	1	106	10.7	1.7	150	33.4%	0.30 [-0.03, 0.63]	2017
Desai 2018	10.8	3.03	394	10.8	2.7	319	20.7%	0.00 [-0.42, 0.42]	2018
Subtotal (95% CI)			**635**			**590**	**100.0%**	**0.15 [-0.05, 0.34]**	

Heterogeneity: $Chi^2 = 1.39$, df = 2 (P = 0.50); $I^2 = 0\%$
Test for overall effect: Z = 1.49 (P = 0.14)

Test for subgroup differences: $Chi^2 = 4.13$, df = 1 (P = 0.04), $I^2 = 75.8\%$



cc3

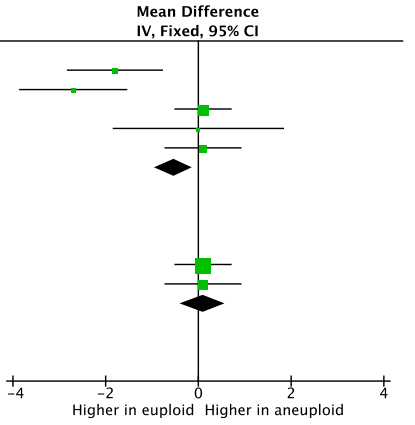
Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
18.1.1 All studies									
Basile 2014	12.2	6.3	361	14	4.8	143	15.3%	-1.80 [-2.82, -0.78]	2014
Chawla 2014	10.8	7	263	13.5	5.6	197	12.0%	-2.70 [-3.85, -1.55]	2014
Mumusoglu 2016	10.4	3.2	242	10.3	3	173	44.0%	0.10 [-0.50, 0.70]	2016
Patel 2016	13.3	6.5	126	13.3	4.7	41	4.7%	0.00 [-1.83, 1.83]	2016
Zhang 2017	13.2	3.4	106	13.1	3.1	150	24.0%	0.10 [-0.72, 0.92]	2017
Subtotal (95% CI)			**1098**			**704**	**100.0%**	**-0.53 [-0.93, -0.13]**	

Heterogeneity: $Chi^2 = 26.40$, df = 4 (P < 0.0001); $I^2 = 85\%$
Test for overall effect: Z = 2.61 (P = 0.009)

18.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	10.4	3.2	242	10.3	3	173	64.7%	0.10 [-0.50, 0.70]	2016
Zhang 2017	13.2	3.4	106	13.1	3.1	150	35.3%	0.10 [-0.72, 0.92]	2017
Subtotal (95% CI)			**348**			**323**	**100.0%**	**0.10 [-0.38, 0.58]**	

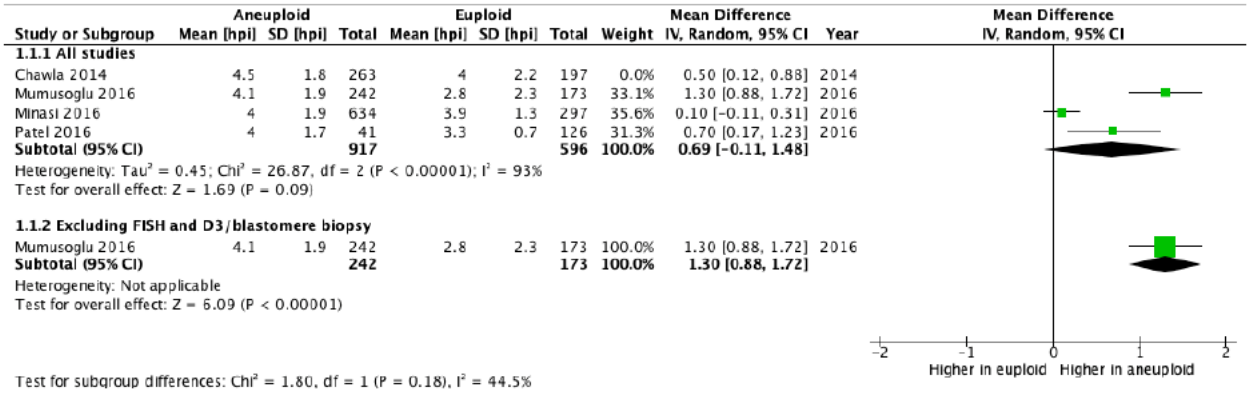
Heterogeneity: $Chi^2 = 0.00$, df = 1 (P = 1.00); $I^2 = 0\%$
Test for overall effect: Z = 0.40 (P = 0.69)

Test for subgroup differences: $Chi^2 = 3.89$, df = 1 (P = 0.05), $I^2 = 74.3\%$

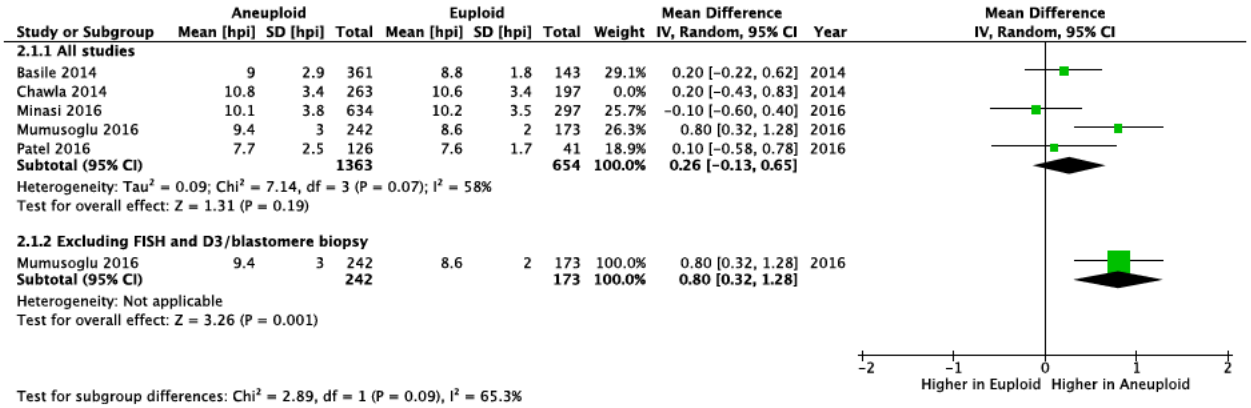


Supplementary Figures 4- Sensitivity analysis for morphokinetics- excluding studies with a high risk of bias according to QUIPS

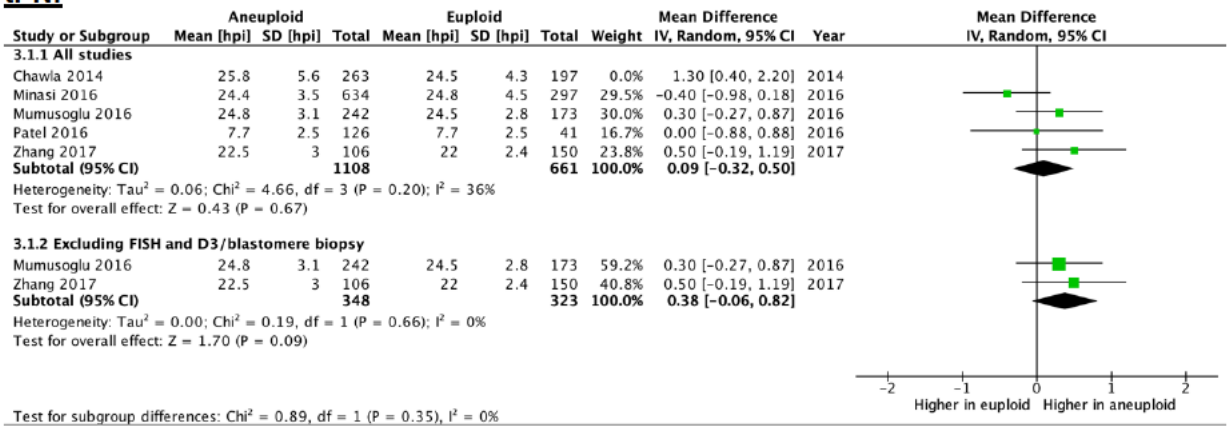
tPB2



tPN

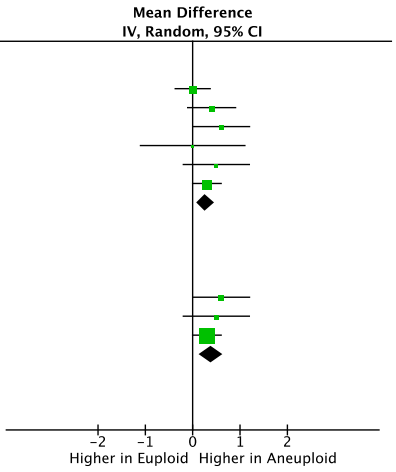


tPNf



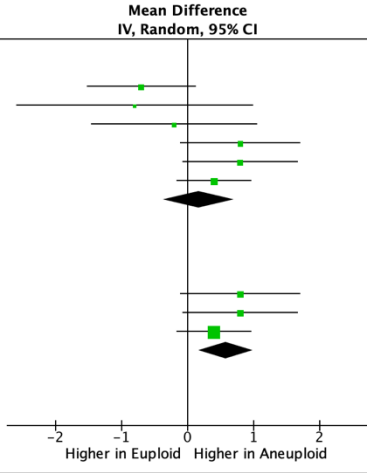
t2

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
4.1.1 All studies									
Chawla 2014	30.6	9.7	263	28.3	7.2	197	0.0%	2.30 [0.76, 3.84]	2014
Basile 2014	26.8	3.4	361	26.8	0.7	143	26.3%	0.00 [−0.37, 0.37]	2014
Minasi 2016	28.4	0.9	634	28	4.4	297	14.0%	0.40 [−0.11, 0.91]	2016
Mumusoglu 2016	27.3	3.3	242	26.7	2.9	173	9.9%	0.60 [0.00, 1.20]	2016
Patel 2016	27.6	3.5	126	27.6	3	41	2.9%	0.00 [−1.10, 1.10]	2016
Zhang 2017	25	3	106	24.5	2.5	150	7.4%	0.50 [−0.20, 1.20]	2017
Desai 2018	26.2	0.3	394	25.9	2.73	319	39.5%	0.30 [−0.00, 0.60]	2018
Subtotal (95% CI)			**1863**			**1123**	**100.0%**	**0.27 [0.08, 0.46]**	
Heterogeneity: Tau² = 0.00; Chi² = 4.16, df = 5 (P = 0.53); I² = 0%									
Test for overall effect: Z = 2.81 (P = 0.005)									
4.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	27.3	3.3	242	26.7	2.9	173	17.5%	0.60 [0.00, 1.20]	2016
Zhang 2017	25	3	106	24.5	2.5	150	13.0%	0.50 [−0.20, 1.20]	2017
Desai 2018	26.2	0.3	394	25.9	2.73	319	69.5%	0.30 [−0.00, 0.60]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.38 [0.13, 0.63]**	
Heterogeneity: Tau² = 0.00; Chi² = 0.90, df = 2 (P = 0.64); I² = 0%									
Test for overall effect: Z = 2.96 (P = 0.003)									
Test for subgroup differences: Chi² = 0.45, df = 1 (P = 0.50), I² = 0%									



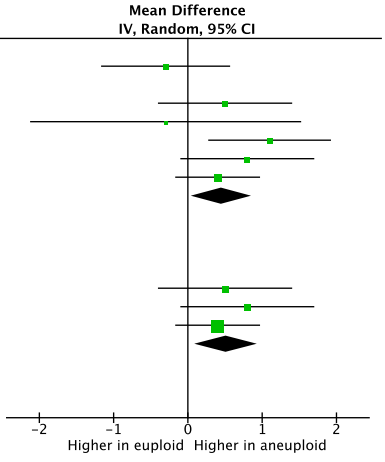
t3

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
5.1.1 All studies									
Chawla 2014	39.7	8.5	263	38.7	7	197	0.0%	1.00 [−0.42, 2.42]	2014
Basile 2014	37.2	4.8	361	37.9	3.96	143	19.3%	−0.70 [−1.52, 0.12]	2014
Patel 2016	37.9	5.2	41	38.7	4.6	126	7.3%	−0.80 [−2.58, 0.98]	2016
Minasi 2016	37.2	8.99	297	37.4	9.23	634	12.1%	−0.20 [−1.45, 1.05]	2016
Mumusoglu 2016	37.6	4.5	242	36.8	4.7	173	17.6%	0.80 [−0.10, 1.70]	2016
Zhang 2017	36	3.6	106	35.2	3.3	150	18.3%	0.80 [−0.07, 1.67]	2017
Desai 2018	37.1	4	394	36.7	3.6	319	25.3%	0.40 [−0.16, 0.96]	2018
Subtotal (95% CI)			**1441**			**1545**	**100.0%**	**0.17 [−0.37, 0.71]**	
Heterogeneity: Tau² = 0.22; Chi² = 10.28, df = 5 (P = 0.07); I² = 51%									
Test for overall effect: Z = 0.62 (P = 0.54)									
5.1.2 Exclude FISH and D3/blastomere biopsy									
Mumusoglu 2016	37.6	4.5	242	36.8	4.7	173	21.3%	0.80 [−0.10, 1.70]	2016
Zhang 2017	36	3.6	106	35.2	3.3	150	23.1%	0.80 [−0.07, 1.67]	2017
Desai 2018	37.1	4	394	36.7	3.6	319	55.5%	0.40 [−0.16, 0.96]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.58 [0.16, 0.99]**	
Heterogeneity: Tau² = 0.00; Chi² = 0.88, df = 2 (P = 0.65); I² = 0%									
Test for overall effect: Z = 2.72 (P = 0.007)									
Test for subgroup differences: Chi² = 1.37, df = 1 (P = 0.24), I² = 26.7%									



t4

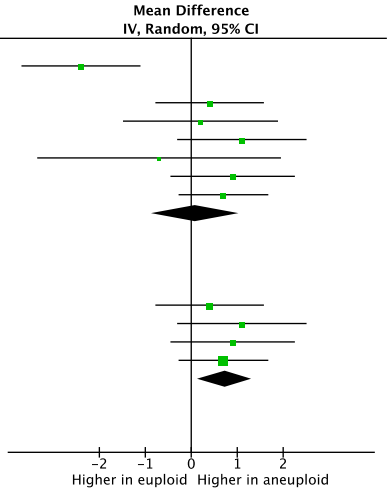
Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
6.1.1 All studies									
Basile 2014	39.6	4.84	361	39.9	4.27	143	16.7%	−0.30 [−1.16, 0.56]	2014
Chawla 2014	41.5	8.2	263	40.5	7.2	197	0.0%	1.00 [−0.41, 2.41]	2014
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	15.7%	0.50 [−0.39, 1.39]	2016
Patel 2016	40.1	5	126	40.4	5.2	41	4.6%	−0.30 [−2.12, 1.52]	2016
Minasi 2016	41.1	6.4	634	40	5.7	297	18.0%	1.10 [0.28, 1.92]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	15.7%	0.80 [−0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	29.4%	0.40 [−0.16, 0.96]	2018
Subtotal (95% CI)			**1863**			**1123**	**100.0%**	**0.46 [0.05, 0.86]**	
Heterogeneity: Tau² = 0.06; Chi² = 6.64, df = 5 (P = 0.25); I² = 25%									
Test for overall effect: Z = 2.21 (P = 0.03)									
6.1.2 Excluded FISH and D3/ blastomere biopsy									
Mumusoglu 2016	39.3	4.7	242	38.8	4.5	173	22.0%	0.50 [−0.39, 1.39]	2016
Zhang 2017	37.1	3.9	106	36.3	3.1	150	22.1%	0.80 [−0.09, 1.69]	2017
Desai 2018	38.5	4	394	38.1	3.64	319	55.9%	0.40 [−0.16, 0.96]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.51 [0.09, 0.93]**	
Heterogeneity: Tau² = 0.00; Chi² = 0.55, df = 2 (P = 0.76); I² = 0%									
Test for overall effect: Z = 2.38 (P = 0.02)									
Test for subgroup differences: Chi² = 0.03, df = 1 (P = 0.85), I² = 0%									



t5

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
7.1.1 All studies									
Basile 2014	49.4	7.75	361	51.8	6.1	143	15.4%	−2.40 [−3.68, −1.12]	2014
Chawla 2014	50.1	9.6	263	52.3	8.6	197	0.0%	−2.20 [−3.87, −0.53]	2014
Yang 2014	50.5	4.7	135	50.1	4.8	121	16.2%	0.40 [−0.77, 1.57]	2014
Minasi 2016	50.6	12.8	634	50.4	11.8	297	12.9%	0.20 [−1.47, 1.87]	2016
Mumusoglu 2016	49.7	7.2	242	48.6	7.1	173	14.7%	1.10 [−0.29, 2.49]	2016
Patel 2016	51.3	8.3	126	52	7.2	41	8.1%	−0.70 [−3.34, 1.94]	2016
Zhang 2017	49.2	5.3	106	48.3	5.5	150	15.1%	0.90 [−0.44, 2.24]	2017
Desai 2018	49.9	6.6	319	49.2	6.4	394	17.6%	0.70 [−0.26, 1.66]	2018
Subtotal (95% CI)			**1923**			**1319**	**100.0%**	**0.08 [−0.87, 1.04]**	
Heterogeneity: Tau² = 1.11; Chi² = 20.07, df = 6 (P = 0.003); I² = 70%									
Test for overall effect: Z = 0.17 (P = 0.86)									
7.1.2 Exclude FISH and D3/blastomere biopsy									
Yang 2014	50.5	4.7	135	50.1	4.8	121	25.4%	0.40 [−0.77, 1.57]	2014
Mumusoglu 2016	49.7	7.2	242	48.6	7.1	173	17.8%	1.10 [−0.29, 2.49]	2016
Zhang 2017	49.2	5.3	106	48.3	5.5	150	19.3%	0.90 [−0.44, 2.24]	2017
Desai 2018	49.9	6.6	319	49.2	6.4	394	37.4%	0.70 [−0.26, 1.66]	2018
Subtotal (95% CI)			**802**			**838**	**100.0%**	**0.73 [0.15, 1.32]**	
Heterogeneity: Tau² = 0.00; Chi² = 0.64, df = 3 (P = 0.89); I² = 0%									
Test for overall effect: Z = 2.44 (P = 0.01)									

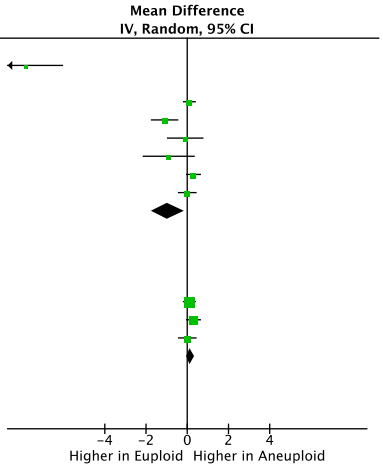
Test for subgroup differences: Chi² = 1.29, df = 1 (P = 0.26), I² = 22.6%



cc2

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
17.1.1 All studies									
Chavez 2012	4	5.2	37	11.8	0.71	8	9.3%	−7.80 [−9.55, −6.05]	2012
Chawla 2014	9.1	4.9	263	10.5	4.2	197	0.0%	−1.40 [−2.23, −0.57]	2014
Yang 2014	11.3	1.1	135	11.2	1.2	121	16.6%	0.10 [−0.18, 0.38]	2014
Basile 2014	10.4	4.4	361	11.5	2.7	143	15.3%	−1.10 [−1.73, −0.47]	2014
Minasi 2016	10.4	6.4	634	10.5	6.1	297	14.2%	−0.10 [−0.95, 0.75]	2016
Patel 2016	10.2	3.8	126	11.1	3.4	41	12.0%	−0.90 [−2.13, 0.33]	2016
Zhang 2017	11	1	106	10.7	1.7	150	16.4%	0.30 [−0.03, 0.63]	2017
Desai 2018	10.8	3.03	394	10.8	2.7	319	16.2%	0.00 [−0.42, 0.42]	2018
Subtotal (95% CI)			**1793**			**1079**	**100.0%**	**−0.95 [−1.75, −0.16]**	
Heterogeneity: Tau² = 0.97; Chi² = 93.41, df = 6 (P < 0.00001); I² = 94%									
Test for overall effect: Z = 2.34 (P = 0.02)									
17.1.2 Excluding FISH and D3/blastomere biopsy									
Yang 2014	11.3	1.1	135	11.2	1.2	121	45.9%	0.10 [−0.18, 0.38]	2014
Zhang 2017	11	1	106	10.7	1.7	150	33.4%	0.30 [−0.03, 0.63]	2017
Desai 2018	10.8	3.03	394	10.8	2.7	319	20.7%	0.00 [−0.42, 0.42]	2018
Subtotal (95% CI)			**635**			**590**	**100.0%**	**0.15 [−0.05, 0.34]**	
Heterogeneity: Tau² = 0.00; Chi² = 1.39, df = 2 (P = 0.50); I² = 0%									
Test for overall effect: Z = 1.49 (P = 0.14)									

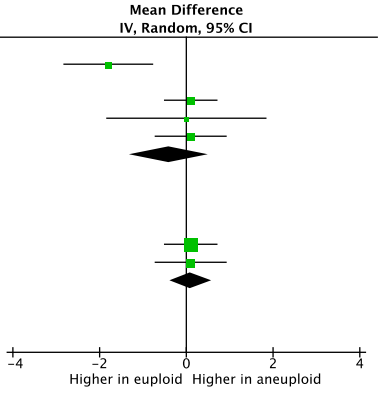
Test for subgroup differences: Chi² = 6.91, df = 1 (P = 0.009), I² = 85.5%



cc3

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
18.1.1 All studies									
Basile 2014	12.2	6.3	361	14	4.8	143	25.1%	−1.80 [−2.82, −0.78]	2014
Chawla 2014	10.8	7	263	13.5	5.6	197	0.0%	−2.70 [−3.85, −1.55]	2014
Mumusoglu 2016	10.4	3.2	242	10.3	3	173	31.7%	0.10 [−0.50, 0.70]	2016
Patel 2016	13.3	6.5	126	13.3	4.7	41	14.7%	0.00 [−1.83, 1.83]	2016
Zhang 2017	13.2	3.4	106	13.1	3.1	150	28.4%	0.10 [−0.72, 0.92]	2017
Subtotal (95% CI)			**835**			**507**	**100.0%**	**−0.39 [−1.30, 0.52]**	
Heterogeneity: Tau² = 0.58; Chi² = 10.94, df = 3 (P = 0.01); I² = 73%									
Test for overall effect: Z = 0.85 (P = 0.40)									
18.1.2 Excluding FISH and D3/blastomere biopsy									
Mumusoglu 2016	10.4	3.2	242	10.3	3	173	64.7%	0.10 [−0.50, 0.70]	2016
Zhang 2017	13.2	3.4	106	13.1	3.1	150	35.3%	0.10 [−0.72, 0.92]	2017
Subtotal (95% CI)			**348**			**323**	**100.0%**	**0.10 [−0.38, 0.58]**	
Heterogeneity: Tau² = 0.00; Chi² = 0.00, df = 1 (P = 1.00); I² = 0%									
Test for overall effect: Z = 0.40 (P = 0.69)									

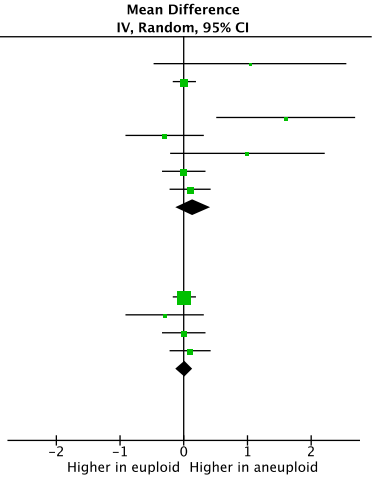
Test for subgroup differences: Chi² = 0.88, df = 1 (P = 0.35), I² = 0%



S2

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
19.1.1 All studies									
Chavez 2012	2	4.3	37	0.96	0.84	8	3.0%	1.04 [-0.46, 2.54]	2012
Yang 2014	0.78	0.71	135	0.77	0.69	121	29.4%	0.01 [-0.16, 0.18]	2014
Chawla 2014	2.1	4	263	1.8	3.3	197	0.0%	0.30 [-0.37, 0.97]	2014
Minasi 2016	4.2	7.7	634	2.6	7.9	297	5.3%	1.60 [0.52, 2.68]	2016
Mumusoglu 2016	1.8	3.1	242	2.1	3.1	173	12.5%	-0.30 [-0.90, 0.30]	2016
Patel 2016	2.3	4	126	1.3	3.2	41	4.4%	1.00 [-0.20, 2.20]	2016
Zhang 2017	1.1	1.2	106	1.1	1.5	150	22.3%	0.00 [-0.33, 0.33]	2017
Desai 2018	1.5	1	394	1.4	2.7	319	23.1%	0.10 [-0.21, 0.41]	2018
Subtotal (95% CI)			**1674**			**1109**	**100.0%**	**0.15 [-0.12, 0.42]**	
Heterogeneity: $Tau^2 = 0.06$; $Chi^2 = 13.65$, df = 6 (P = 0.03); $I^2 = 56\%$									
Test for overall effect: Z = 1.07 (P = 0.28)									
19.1.2 Excluding FISH									
Yang 2014	0.78	0.71	135	0.77	0.69	121	60.6%	0.01 [-0.16, 0.18]	2014
Mumusoglu 2016	1.8	3.1	242	2.1	3.1	173	4.9%	-0.30 [-0.90, 0.30]	2016
Zhang 2017	1.1	1.2	106	1.1	1.5	150	16.3%	0.00 [-0.33, 0.33]	2017
Desai 2018	1.5	1	394	1.4	2.7	319	18.3%	0.10 [-0.21, 0.41]	2018
Subtotal (95% CI)			**877**			**763**	**100.0%**	**0.01 [-0.12, 0.14]**	
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 1.33$, df = 3 (P = 0.72); $I^2 = 0\%$									
Test for overall effect: Z = 0.14 (P = 0.89)									

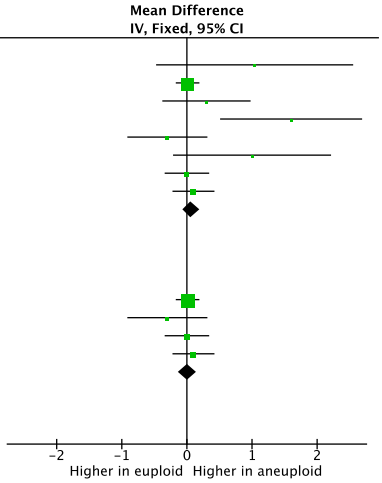
Test for subgroup differences: $Chi^2 = 0.81$, df = 1 (P = 0.37), $I^2 = 0\%$



S2

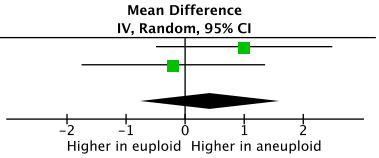
Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
19.1.1 All studies									
Chavez 2012	2	4.3	37	0.96	0.84	8	0.7%	1.04 [-0.46, 2.54]	2012
Yang 2014	0.78	0.71	135	0.77	0.69	121	56.3%	0.01 [-0.16, 0.18]	2014
Chawla 2014	2.1	4	263	1.8	3.3	197	3.7%	0.30 [-0.37, 0.97]	2014
Minasi 2016	4.2	7.7	634	2.6	7.9	297	1.4%	1.60 [0.52, 2.68]	2016
Mumusoglu 2016	1.8	3.1	242	2.1	3.1	173	4.5%	-0.30 [-0.90, 0.30]	2016
Patel 2016	2.3	4	126	1.3	3.2	41	1.1%	1.00 [-0.20, 2.20]	2016
Zhang 2017	1.1	1.2	106	1.1	1.5	150	15.1%	0.00 [-0.33, 0.33]	2017
Desai 2018	1.5	1	394	1.4	2.7	319	17.0%	0.10 [-0.21, 0.41]	2018
Subtotal (95% CI)			**1937**			**1306**	**100.0%**	**0.06 [-0.07, 0.19]**	
Heterogeneity: $Chi^2 = 14.16$, df = 7 (P = 0.05); $I^2 = 51\%$									
Test for overall effect: Z = 0.94 (P = 0.34)									
19.1.2 Excluding FISH									
Yang 2014	0.78	0.71	135	0.77	0.69	121	60.6%	0.01 [-0.16, 0.18]	2014
Mumusoglu 2016	1.8	3.1	242	2.1	3.1	173	4.9%	-0.30 [-0.90, 0.30]	2016
Zhang 2017	1.1	1.2	106	1.1	1.5	150	16.3%	0.00 [-0.33, 0.33]	2017
Desai 2018	1.5	1	394	1.4	2.7	319	18.3%	0.10 [-0.21, 0.41]	2018
Subtotal (95% CI)			**877**			**763**	**100.0%**	**0.01 [-0.12, 0.14]**	
Heterogeneity: $Chi^2 = 1.33$, df = 3 (P = 0.72); $I^2 = 0\%$									
Test for overall effect: Z = 0.14 (P = 0.89)									

Test for subgroup differences: $Chi^2 = 0.31$, df = 1 (P = 0.58), $I^2 = 0\%$



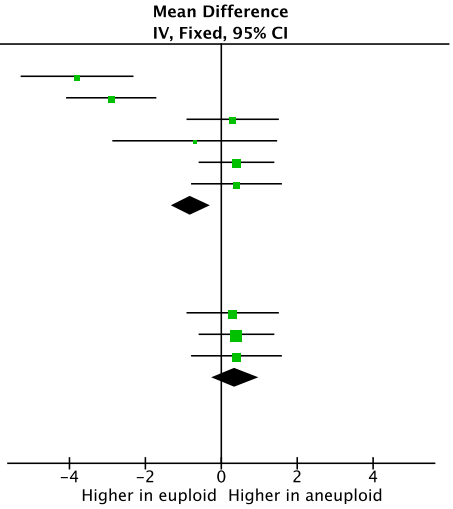
S3

Study or Subgroup	Aneuploid Mean [hpi]	Aneuploid SD [hpi]	Aneuploid Total	Euploid Mean [hpi]	Euploid SD [hpi]	Euploid Total	Weight	Mean Difference IV, Random, 95% CI	Year
Mumusoglu 2016	9.7	8.1	242	8.7	7.2	173	51.7%	1.00 [-0.48, 2.48]	2016
Zhang 2017	6.5	5.9	106	6.7	6.6	150	48.3%	-0.20 [-1.74, 1.34]	2017
Total (95% CI)			**348**			**323**	**100.0%**	**0.42 [-0.76, 1.60]**	
Heterogeneity: $Tau^2 = 0.13$; $Chi^2 = 1.21$, df = 1 (P = 0.27); $I^2 = 17\%$									
Test for overall effect: Z = 0.70 (P = 0.48)									



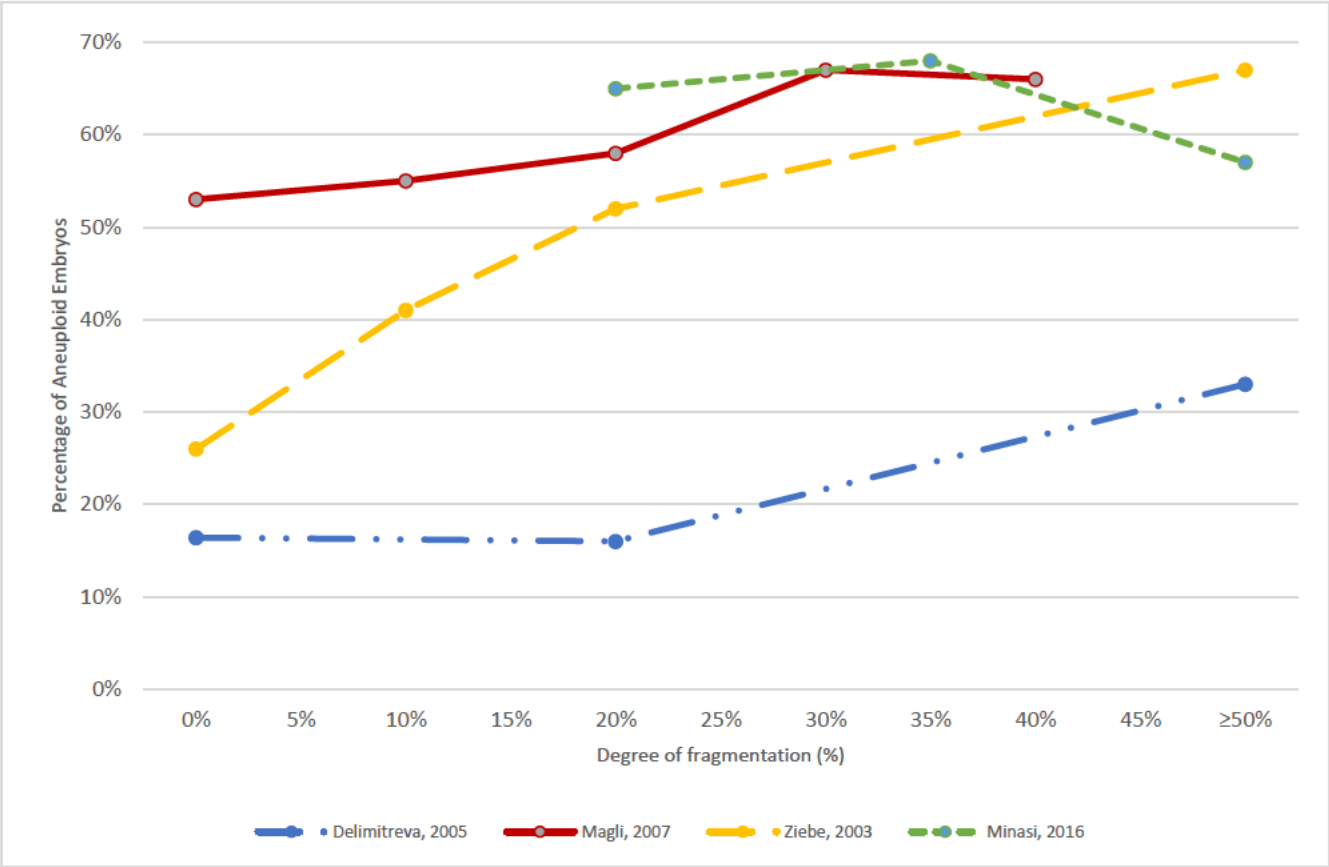
t5-t2

Study or Subgroup	Aneuploid Mean	SD	Total	Euploid Mean	SD	Total	Weight	Mean Difference IV, Fixed, 95% CI	Year
21.1.1 All studies									
Chawla 2014	20.1	8.6	263	23.9	7.4	197	12.0%	-3.80 [-5.27, -2.33]	2014
Basile 2014	22.6	7.2	361	25.5	5.5	143	18.9%	-2.90 [-4.07, -1.73]	2014
Mumusoglu 2016	22.5	6.3	242	22.2	6	173	18.0%	0.30 [-0.90, 1.50]	2016
Patel 2016	23.7	7.2	126	24.4	5.7	41	5.6%	-0.70 [-2.85, 1.45]	2016
Zhang 2017	24.2	3.8	106	23.8	4.1	150	27.0%	0.40 [-0.58, 1.38]	2017
Desai 2018	23.7	5.5	394	23.3	9.5	319	18.6%	0.40 [-0.78, 1.58]	2018
Subtotal (95% CI)			**1492**			**1023**	**100.0%**	**-0.80 [-1.31, -0.30]**	
Heterogeneity: $Chi^2 = 41.58$, df = 5 ($P < 0.00001$); $I^2 = 88\%$									
Test for overall effect: Z = 3.11 (P = 0.002)									
21.1.2 Excluding FISH and blastomere biopsy									
Mumusoglu 2016	22.5	6.3	242	22.2	6	173	28.3%	0.30 [-0.90, 1.50]	2016
Zhang 2017	24.2	3.8	106	23.8	4.1	150	42.4%	0.40 [-0.58, 1.38]	2017
Desai 2018	23.7	5.5	394	23.3	9.5	319	29.3%	0.40 [-0.78, 1.58]	2018
Subtotal (95% CI)			**742**			**642**	**100.0%**	**0.37 [-0.26, 1.01]**	
Heterogeneity: $Chi^2 = 0.02$, df = 2 ($P = 0.99$); $I^2 = 0\%$									
Test for overall effect: Z = 1.15 (P = 0.25)									

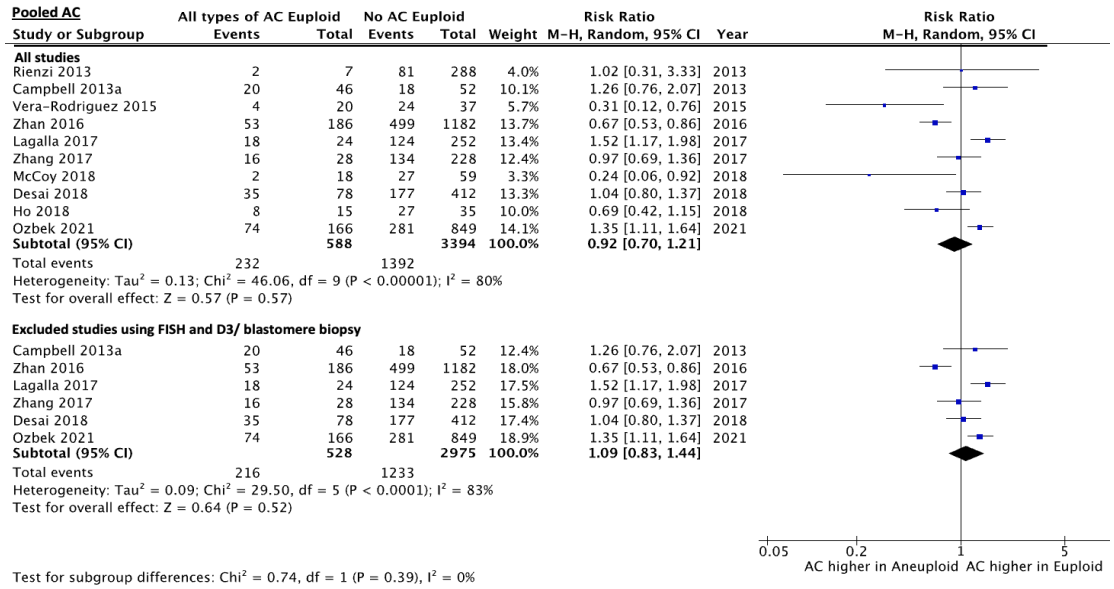


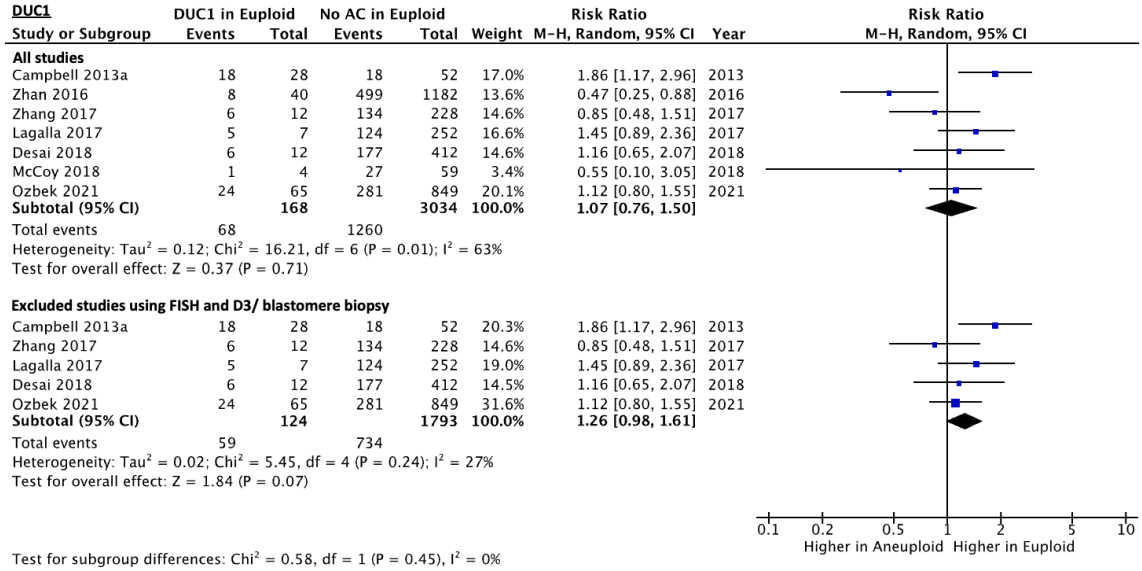
Test for subgroup differences: $Chi^2 = 8.03$, df = 1 ($P = 0.005$), $I^2 = 87.5\%$

Supplementary Figure 5 - Line graph displaying the relationship between fragmentation (%) and prevalence of aneuploidy (4 out of 8 studies included)

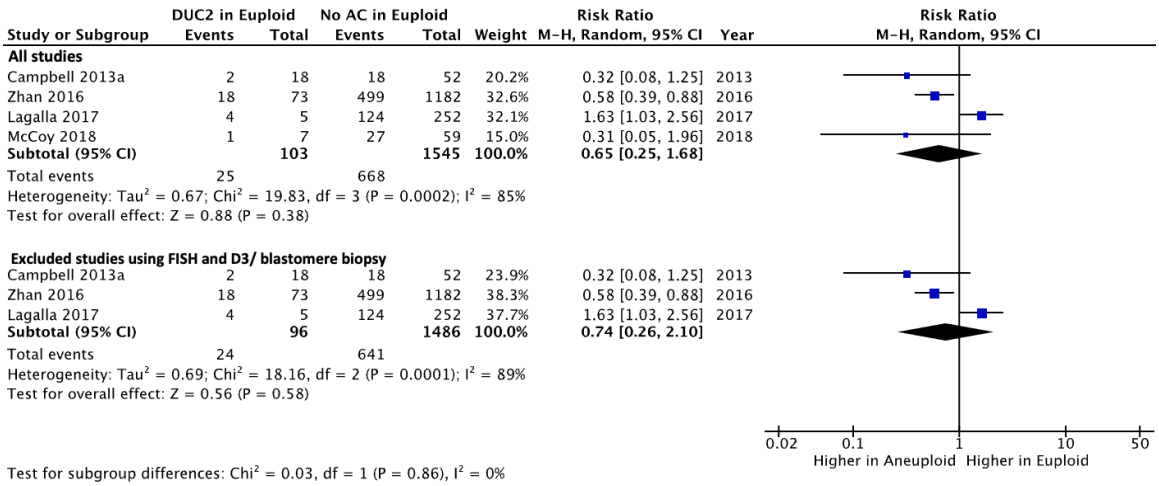


Supplementary Figures 6- Relative risk of euploidy embryos displaying pooled abnormal cleavage, DUC1, DUC2, respectively





DUC2



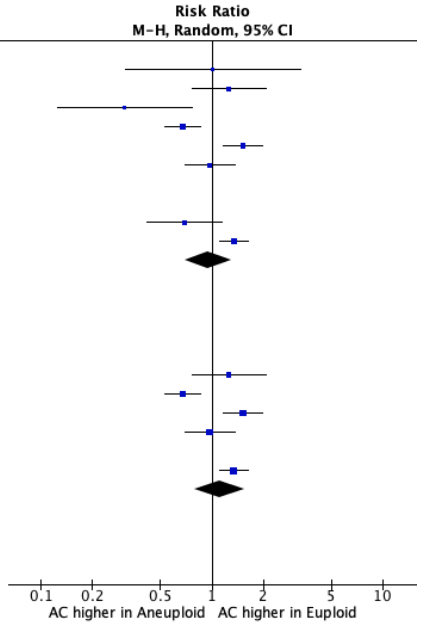
Test for subgroup differences: $Chi^2 = 0.03$, df = 1 (P = 0.86), $I^2 = 0\%$

Supplementary Figures 7- Sensitivity analysis for morphology- excluding studies with a high risk of bias according to QUIPS

Pooled Abnormal Cleavage

Study or Subgroup	All types of AC Euploid Events	Total	No AC Euploid Events	Total	Weight	Risk Ratio M-H, Random, 95% CI	Year
22.1.1 All studies							
Rienzi 2013	2	7	81	288	5.1%	1.02 [0.31, 3.33]	2013
Campbell 2013a	20	46	18	52	12.3%	1.26 [0.76, 2.07]	2013
Vera-Rodriguez 2015	4	20	24	37	7.2%	0.31 [0.12, 0.76]	2015
Zhan 2016	53	186	499	1182	16.1%	0.67 [0.53, 0.86]	2016
Lagalla 2017	18	24	124	252	15.8%	1.52 [1.17, 1.98]	2017
Zhang 2017	16	28	134	228	14.7%	0.97 [0.69, 1.36]	2017
McCoy 2018	2	18	27	59	0.0%	0.24 [0.06, 0.92]	2018
Desai 2018	35	78	177	412	0.0%	1.04 [0.80, 1.37]	2018
Ho 2018	8	15	27	35	12.2%	0.69 [0.42, 1.15]	2018
Ozbek 2021	74	166	281	849	16.6%	1.35 [1.11, 1.64]	2021
Subtotal (95% CI)		**492**		**2923**	**100.0%**	**0.95 [0.69, 1.30]**	
Total events	195		1188				
Heterogeneity: Tau² = 0.15; Chi² = 40.86, df = 7 (P < 0.00001); I² = 83%							
Test for overall effect: Z = 0.32 (P = 0.75)							
22.1.2 Excluding studies using FISH and D3							
Campbell 2013a	20	46	18	52	15.9%	1.26 [0.76, 2.07]	2013
Zhan 2016	53	186	499	1182	21.5%	0.67 [0.53, 0.86]	2016
Lagalla 2017	18	24	124	252	21.0%	1.52 [1.17, 1.98]	2017
Zhang 2017	16	28	134	228	19.4%	0.97 [0.69, 1.36]	2017
Desai 2018	35	78	177	412	0.0%	1.04 [0.80, 1.37]	2018
Ozbek 2021	74	166	281	849	22.2%	1.35 [1.11, 1.64]	2021
Subtotal (95% CI)		**450**		**2563**	**100.0%**	**1.11 [0.79, 1.56]**	
Total events	181		1056				
Heterogeneity: Tau² = 0.13; Chi² = 29.75, df = 4 (P < 0.00001); I² = 87%							
Test for overall effect: Z = 0.58 (P = 0.56)							

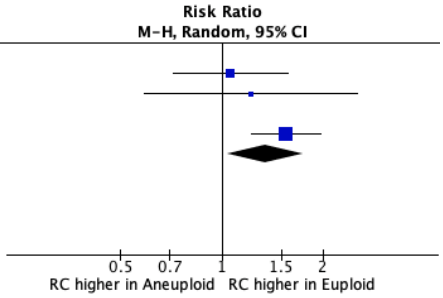
Test for subgroup differences: Chi² = 0.42, df = 1 (P = 0.52), I² = 0%



Reverse Cleavage

Study or Subgroup	RC in Euploid Events	Total	No RC in Euploid Events	Total	Weight	Risk Ratio M-H, Random, 95% CI	Year
23.1.1 All studies							
Zhang 2017	10	16	134	228	30.9%	1.06 [0.72, 1.58]	2017
Lagalla 2017	3	5	124	252	11.3%	1.22 [0.59, 2.52]	2017
Desai 2018	10	19	177	412	0.0%	1.23 [0.79, 1.90]	2018
Ozbek 2021	40	78	281	849	57.9%	1.55 [1.22, 1.96]	2021
Subtotal (95% CI)		**99**		**1329**	**100.0%**	**1.34 [1.04, 1.74]**	
Total events	53		539				
Heterogeneity: Tau² = 0.02; Chi² = 2.71, df = 2 (P = 0.26); I² = 26%							
Test for overall effect: Z = 2.25 (P = 0.02)							

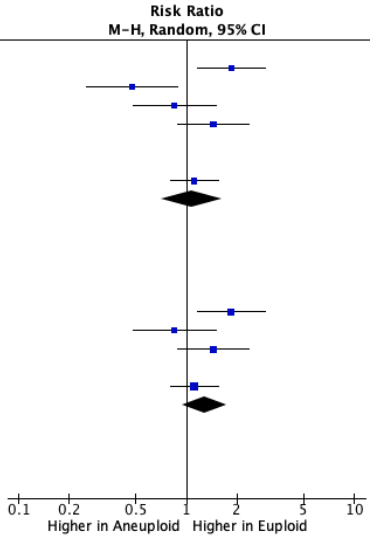
Test for subgroup differences: Not applicable



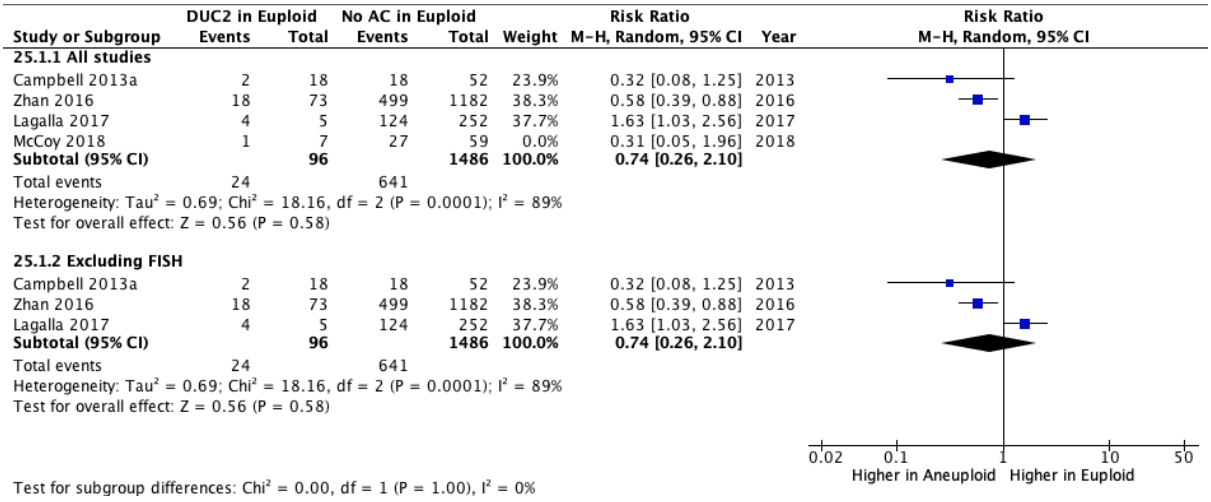
DUC1

Study or Subgroup	DUC1 in Euploid Events	Total	No AC in Euploid Events	Total	Weight	Risk Ratio M-H, Random, 95% CI	Year
24.1.1 All studies							
Campbell 2013a	18	28	18	52	20.7%	1.86 [1.17, 2.96]	2013
Zhan 2016	8	40	499	1182	17.2%	0.47 [0.25, 0.88]	2016
Zhang 2017	6	12	134	228	18.2%	0.85 [0.48, 1.51]	2017
Lagalla 2017	5	7	124	252	20.2%	1.45 [0.89, 2.36]	2017
McCoy 2018	1	4	27	59	0.0%	0.55 [0.10, 3.05]	2018
Desai 2018	6	12	177	412	0.0%	1.16 [0.65, 2.07]	2018
Ozbek 2021	24	65	281	849	23.6%	1.12 [0.80, 1.55]	2021
Subtotal (95% CI)		**152**		**2563**	**100.0%**	**1.07 [0.70, 1.64]**	
Total events	61		1056				
Heterogeneity: Tau² = 0.17; Chi² = 15.50, df = 4 (P = 0.004); I² = 74%							
Test for overall effect: Z = 0.33 (P = 0.74)							
24.1.2 Excluding FISH							
Campbell 2013a	18	28	18	52	24.4%	1.86 [1.17, 2.96]	2013
Zhang 2017	6	12	134	228	18.6%	0.85 [0.48, 1.51]	2017
Lagalla 2017	5	7	124	252	23.2%	1.45 [0.89, 2.36]	2017
Desai 2018	6	12	177	412	0.0%	1.16 [0.65, 2.07]	2018
Ozbek 2021	24	65	281	849	33.8%	1.12 [0.80, 1.55]	2021
Subtotal (95% CI)		**112**		**1381**	**100.0%**	**1.28 [0.94, 1.73]**	
Total events	53		557				
Heterogeneity: Tau² = 0.04; Chi² = 5.39, df = 3 (P = 0.15); I² = 44%							
Test for overall effect: Z = 1.58 (P = 0.12)							

Test for subgroup differences: Chi² = 0.42, df = 1 (P = 0.51), I² = 0%



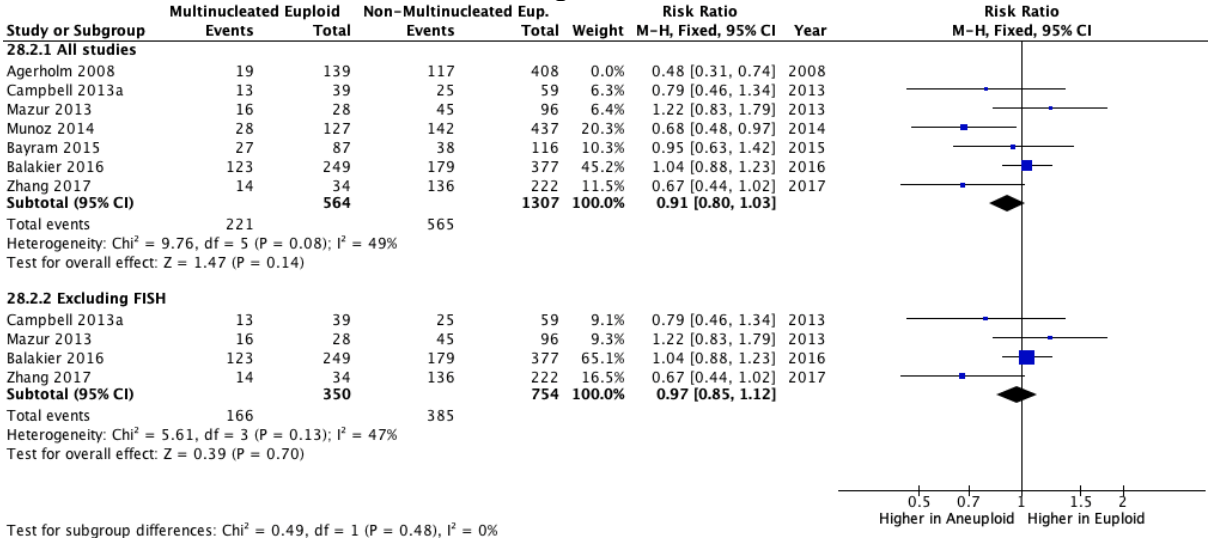
DUC2



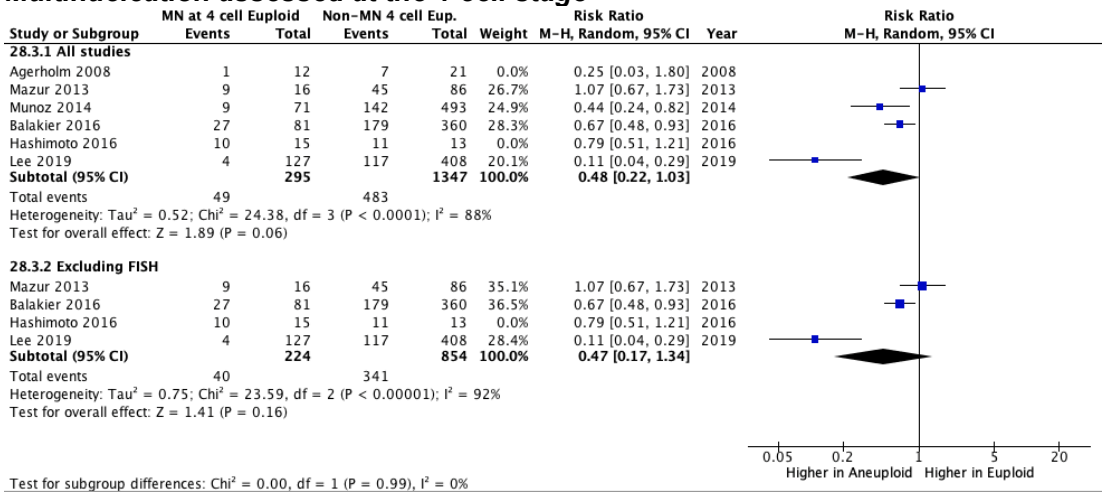
Multinucleation on day 2

All studies included in the analysis have a high risk of bias therefore would be excluded

Multinucleation assessed at the 2-cell stage

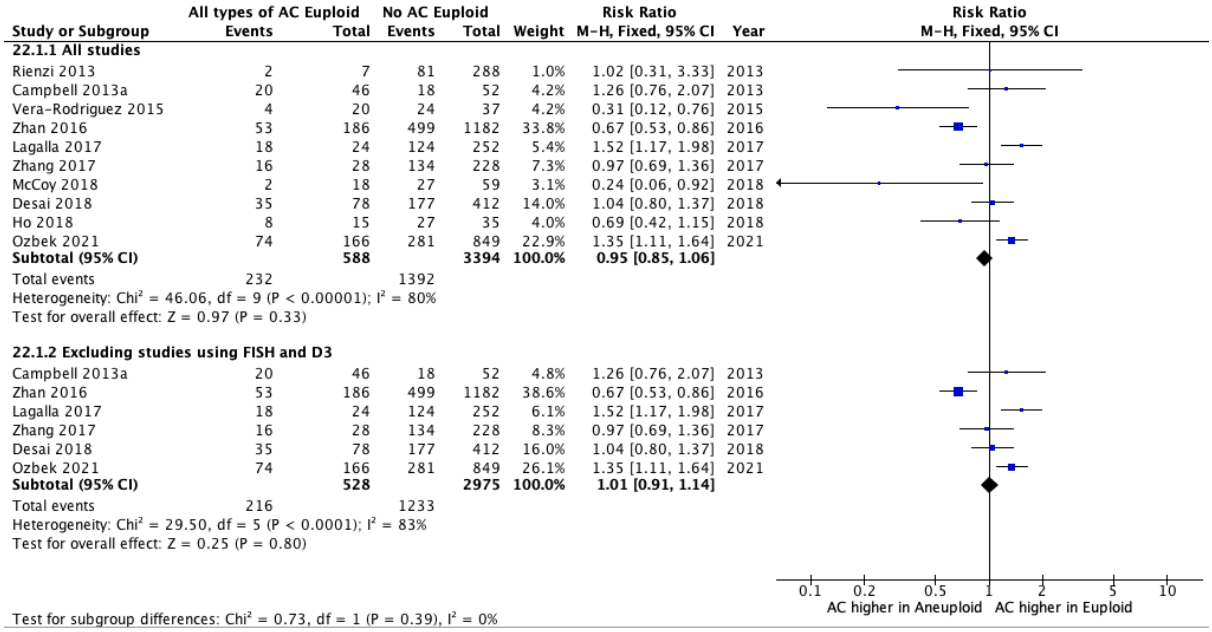


Multinucleation assessed at the 4-cell stage

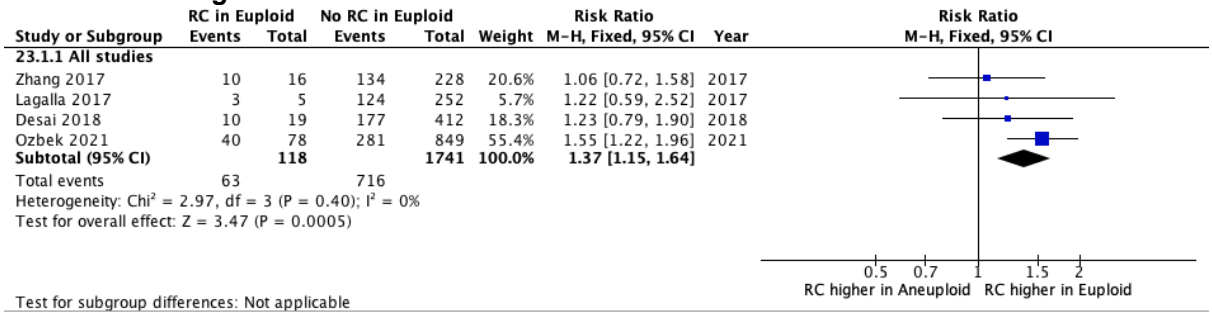


Supplementary Figures 8- Sensitivity analysis for morphology- Fixed effects

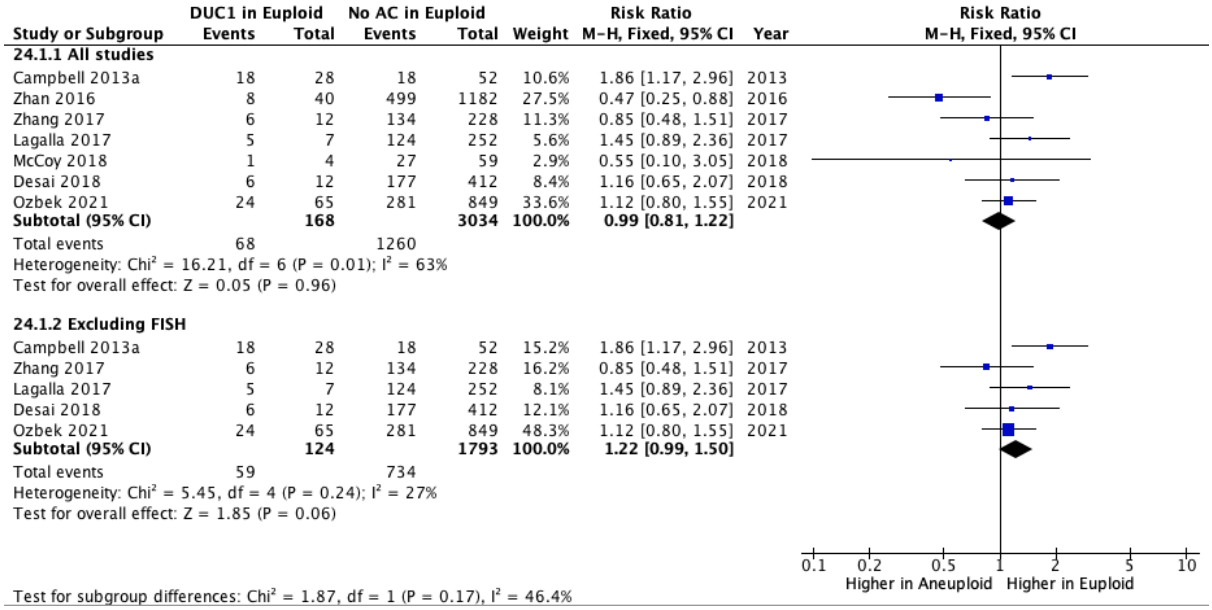
Pooled Abnormal Cleavage



Reverse Cleavage

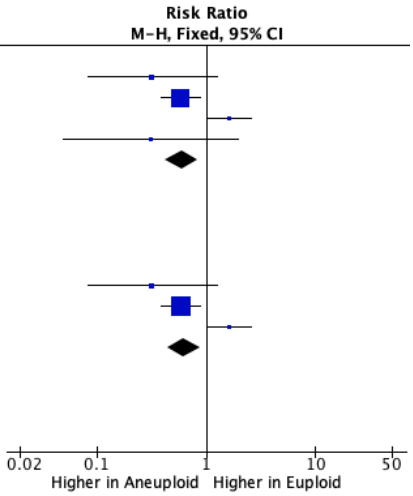


DUC1



DUC2

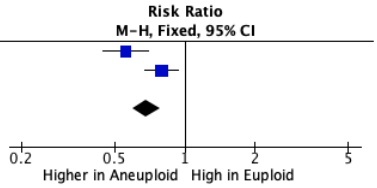
Study or Subgroup	DUC2 in Euploid Events	Total	No AC in Euploid Events	Total	Weight	Risk Ratio M-H, Fixed, 95% CI	Year
25.1.1 All studies							
Campbell 2013a	2	18	18	52	11.9%	0.32 [0.08, 1.25]	2013
Zhan 2016	18	73	499	1182	74.6%	0.58 [0.39, 0.88]	2016
Lagalla 2017	4	5	124	252	6.2%	1.63 [1.03, 2.56]	2017
McCoy 2018	1	7	27	59	7.4%	0.31 [0.05, 1.96]	2018
Subtotal (95% CI)		**103**		**1545**	**100.0%**	**0.60 [0.43, 0.84]**	
Total events	25		668				
Heterogeneity: Chi² = 19.83, df = 3 (P = 0.0002); I² = 85%							
Test for overall effect: Z = 2.97 (P = 0.003)							
25.1.2 Excluding FISH							
Campbell 2013a	2	18	18	52	12.8%	0.32 [0.08, 1.25]	2013
Zhan 2016	18	73	499	1182	80.5%	0.58 [0.39, 0.88]	2016
Lagalla 2017	4	5	124	252	6.7%	1.63 [1.03, 2.56]	2017
Subtotal (95% CI)		**96**		**1486**	**100.0%**	**0.62 [0.44, 0.88]**	
Total events	24		641				
Heterogeneity: Chi² = 18.16, df = 2 (P = 0.0001); I² = 89%							
Test for overall effect: Z = 2.72 (P = 0.007)							



Test for subgroup differences: Chi² = 0.02, df = 1 (P = 0.88), I² = 0%

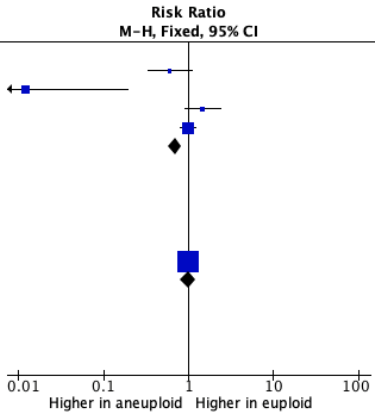
Blastocyst Contraction

Study or Subgroup	Contractions (euploid) Events	Total	No contractions (euploid) Events	Total	Weight	Risk Ratio M-H, Fixed, 95% CI	Year
Gonzalez 2018	98	504	136	392	47.8%	0.56 [0.45, 0.70]	2018
Gazzo 2019	91	196	311	534	52.2%	0.80 [0.67, 0.94]	2019
Total (95% CI)		**700**		**926**	**100.0%**	**0.68 [0.60, 0.78]**	
Total events	189		447				
Heterogeneity: Chi² = 6.29, df = 1 (P = 0.01); I² = 84%							
Test for overall effect: Z = 5.52 (P < 0.00001)							



Multinucleation on day 2

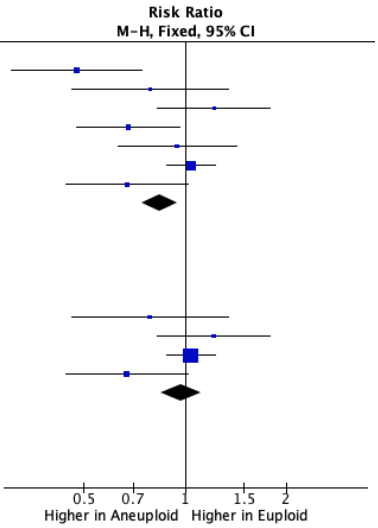
Study or Subgroup	Multinucleated Euploid Events	Total	Non-Multinucleated Eup. Events	Total	Weight	Risk Ratio M-H, Fixed, 95% CI	Year
28.1.1 All studies							
Kligman 1996	8	26	205	403	10.1%	0.60 [0.34, 1.09]	1996
Magli 2001	0	109	518	1367	31.4%	0.01 [0.00, 0.19]	2001
Ambroggio 2011	228	1033	15	99	11.2%	1.46 [0.90, 2.35]	2011
Desai 2018	85	201	177	412	47.3%	0.98 [0.81, 1.20]	2018
Subtotal (95% CI)		**1369**		**2281**	**100.0%**	**0.69 [0.58, 0.83]**	
Total events	321		915				
Heterogeneity: Chi² = 29.98, df = 3 (P < 0.00001); I² = 90%							
Test for overall effect: Z = 3.88 (P = 0.0001)							
28.1.2 Excluding FISH							
Desai 2018	85	201	177	412	100.0%	0.98 [0.81, 1.20]	2018
Subtotal (95% CI)		**201**		**412**	**100.0%**	**0.98 [0.81, 1.20]**	
Total events	85		177				
Heterogeneity: Not applicable							
Test for overall effect: Z = 0.16 (P = 0.87)							



Test for subgroup differences: Chi² = 6.50, df = 1 (P = 0.01), I² = 84.6%

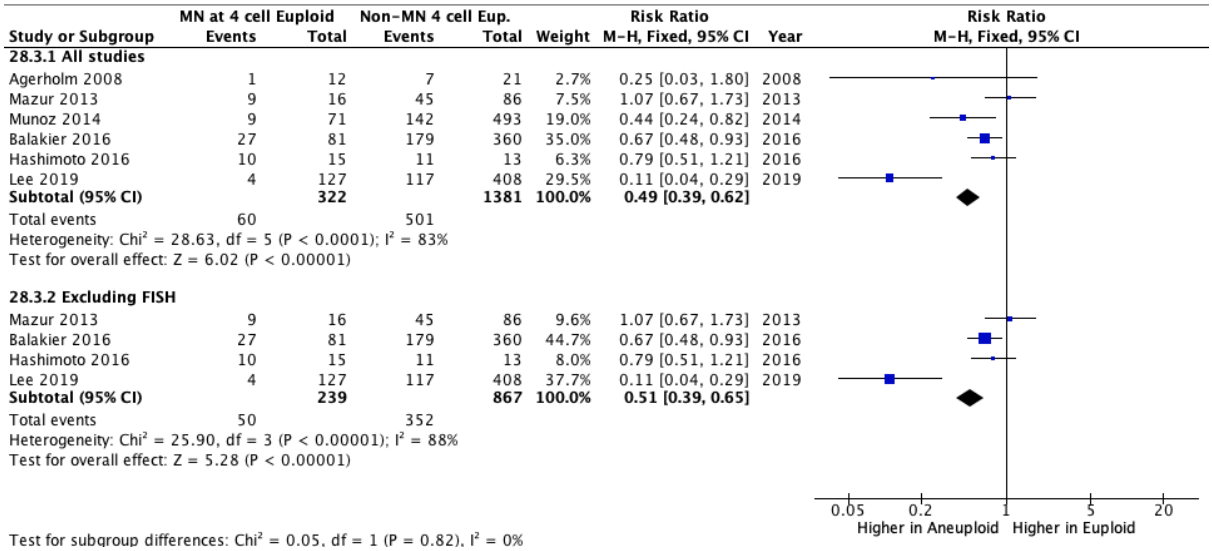
Multinucleation at the 2-cell stage

Study or Subgroup	Multinucleated Euploid Events	Total	Non-Multinucleated Eup. Events	Total	Weight	Risk Ratio M-H, Fixed, 95% CI	Year
28.2.1 All studies							
Agerholm 2008	19	139	117	408	15.9%	0.48 [0.31, 0.74]	2008
Campbell 2013a	13	39	25	59	5.3%	0.79 [0.46, 1.34]	2013
Mazur 2013	16	28	45	96	5.4%	1.22 [0.83, 1.79]	2013
Munoz 2014	28	127	142	437	17.1%	0.68 [0.48, 0.97]	2014
Bayram 2015	27	87	38	116	8.7%	0.95 [0.63, 1.42]	2015
Balakier 2016	123	249	179	377	38.0%	1.04 [0.88, 1.23]	2016
Zhang 2017	14	34	136	222	9.6%	0.67 [0.44, 1.02]	2017
Subtotal (95% CI)		**703**		**1715**	**100.0%**	**0.84 [0.75, 0.95]**	
Total events	240		682				
Heterogeneity: Chi² = 19.13, df = 6 (P = 0.004); I² = 69%							
Test for overall effect: Z = 2.79 (P = 0.005)							
28.2.2 Excluding FISH							
Campbell 2013a	13	39	25	59	9.1%	0.79 [0.46, 1.34]	2013
Mazur 2013	16	28	45	96	9.3%	1.22 [0.83, 1.79]	2013
Balakier 2016	123	249	179	377	65.1%	1.04 [0.88, 1.23]	2016
Zhang 2017	14	34	136	222	16.5%	0.67 [0.44, 1.02]	2017
Subtotal (95% CI)		**350**		**754**	**100.0%**	**0.97 [0.85, 1.12]**	
Total events	166		385				
Heterogeneity: Chi² = 5.61, df = 3 (P = 0.13); I² = 47%							
Test for overall effect: Z = 0.39 (P = 0.70)							

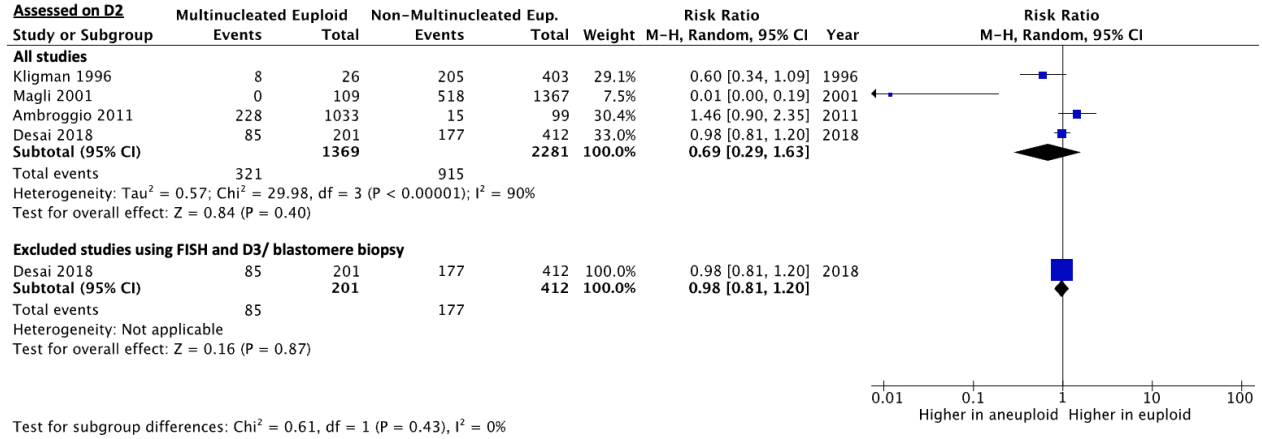


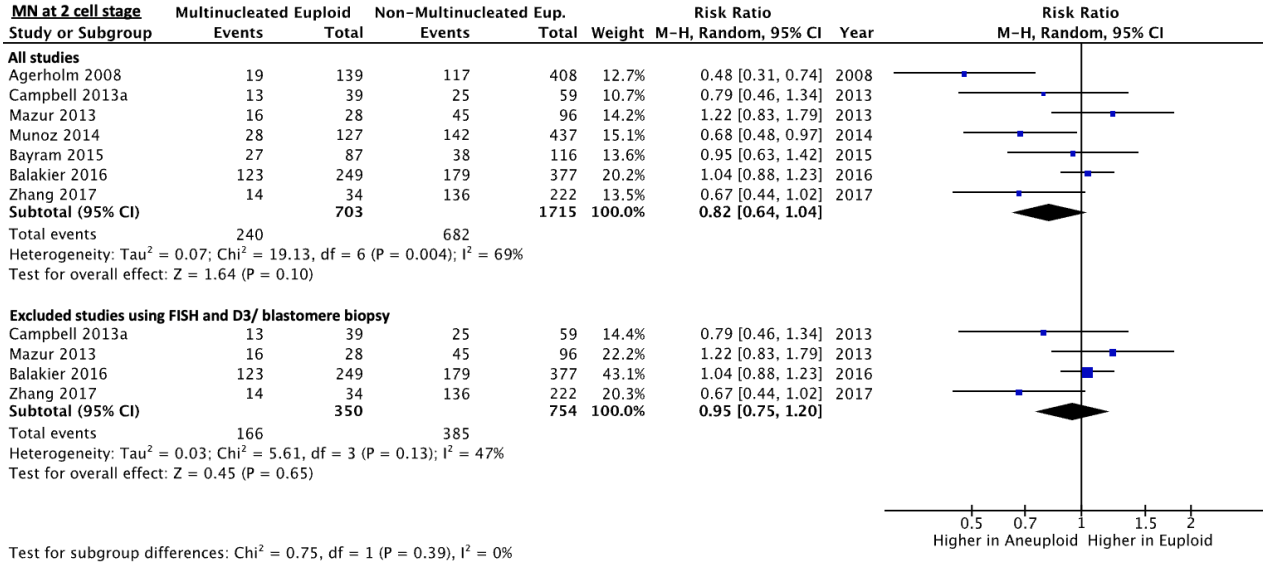
Test for subgroup differences: Chi² = 2.39, df = 1 (P = 0.12), I² = 58.2%

Multinucleation at the 4-cell stage



Supplementary Figures 9- Relative risk of a multinucleated embryo being euploid when assessed on day 2, at the 2 cell stage, respectively.





APPENDIX II- SUPPLEMENTARY DATA FOR CHAPTER 3

Appendix I- IRB Committee Ethical Approval

9th February 2023

Dear Dr Bamford,

Re: IRB Committee ethical approval

- **Full title:** Predicting euploidy for embryos in reproductive medicine- part I and II
- **Sponsor:** CARE fertility and Tommy's National Centre for Miscarriage Research
- **Applicant Institution:** CARE Fertility

On behalf of the Committee, I am pleased to confirm a favourable ethics opinion for the above research on the basis described in the protocol

Approved trial documentation

PREFER I and II- Predicting Euploidy For Embryos in Reproductive Medicine

Version 5.0, dated 8th Feb 2023

IRB Committee composition

Name	Job Title	IRB Committee Role
Mr Ektoras Georgiou	Subspecialty Registrar in Reproductive Medicine and Surgery	Chair
Mrs Oonagh Pickering	Lead Research Nurse, Tommy's National Centre for Miscarriage Research	Member
Janine O'Rourke	Cardiac Network Delivery Lead	Member

Yours faithfully,

Mr Ektoras Georgiou
MB BS, PhD, MRCOG

Supplementary Table 1- TRIPOD Checklist

Title	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	✓
Abstract	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	✓
Background and objectives	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	✓
	Specify the objectives, including whether the study describes the development or validation of the model or both.	✓
Source of data	Describe the study design or source of data (e.g. randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	✓
	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	✓
Participants	Specify key elements of the study setting (e.g. primary care, secondary care, general population) including number and location of centres.	✓
	Describe eligibility criteria for participants.	✓
	Give details of treatments received, if relevant.	✓
Outcome	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	✓
	Report any actions to blind assessment of the outcome to be predicted.	✓
Predictors	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	✓
	Report any actions to blind assessment of predictors for the outcome and other predictors.	✓
Sample size	Explain how the study size was arrived at.	✓
Missing data	Describe how missing data were handled (e.g. complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	✓
Statistical analysis methods	Describe how predictors were handled in the analyses.	✓
	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	✓
	For validation, describe how the predictions were calculated.	✓
	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	✓
	Describe any model updating (e.g., recalibration) arising from the validation, if done.	✓
Risk groups	Provide details on how risk groups were created, if done.	✓
Development vs validation	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	✓
Participants	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	✓
	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	✓
	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	✓
Model development	Specify the number of participants and outcome events in each analysis.	✓
	If done, report the unadjusted association between each candidate predictor and outcome.	✓
Interpretation	For validation, discuss the results with reference to performance in the development data, and any other validation data.	✓
	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	✓
Implications	Discuss the potential clinical use of the model and implications for future research.	✓
Supplementary information	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	✓
Funding	Give the source of funding and the role of the funders for the present study.	✓

Supplementary Table 2- Terminology definitions for AI modelling

Term	Definition
Random forest classifier	RFC is a classification algorithm made up of many decision trees using a bootstrapped dataset. Decision trees create a model by evaluating true or false questions, used for classification. Random forest is an ensemble learning algorithm, meaning that it combines multiple machine learning methods for classification. RFC use a technique called bagging to build the trees in parallel, the forest chooses the classification with the majority of the 'votes' given by the individual trees.
Bootstrapped	A resampling technique where the data is repeatedly drawn from the same sample with replacement.
Extreme gradient boosting	XGBoost has similarities to RFC; however, the trees are added sequentially and fit to correct the prediction errors made by prior models. Therefore, they use previously weak models in order to generate a collectively stronger model.
Deep learning	A type of artificial neural network which is inspired by the biological neurons of the brain. This is comprised of node layers, containing an input layer, one or more hidden layers (where complex non-linear functions are applied with different weights), and an output layer. A dense or fully connected layer is composed of different nodes or units which are controlled by weights, each node has a bias which can be used to shift the output of the nodes. For the first layer, the node gets its input from the data being fed into the network (each data point is connected to each node). The biases and weights are learnt in training to improve accuracy. A series of hidden layers are located between the input and output in which the function applies weights to the input and directs them through the network. After applying the weight and bias of all input neuros they are summed together into a single number. This is then passed through an activation function, in our case a sigmoid function as it is a binary outcome. Each node is fully connected to the nodes in the layers before and after it. There is therefore no restriction on how information can flow through the network and therefore there is a lack of structure.
Hyperparameters	Model parameters are determined during training but there are also adjustable parameters that can be tuned to obtain a model with improved performance. For the RFC algorithm, the following hyperparameters were optimized using the validation set: number of decision trees, criteria to split on each node, maximum tree depth, and number of random features to include at each node for splitting. For XGBoost, this study tuned for: learning rate, maximum depth, number of trees (estimators), the fraction of columns and observations to be randomly sampled for in each tree, and regularization parameters. For deep learning, the keras tuner was used, tuning for epoch number (the number of times that the learning algorithm will work through the training dataset), optimizer (a function that modifies weights), learning rate, and optimal architecture.
Learning rate	A hyperparameter used to govern the pace at which an algorithm updates or learn the values of an estimate, i.e. how often it refreshes what it has learnt in respect to the loss gradient (penalty for bad prediction).
One hot encoding	A method to convert categorical data with multiple categories into variables with only binary vectors
K-fold cross-validation	This is a procedure similar to IECV where the dataset is split into k groups, the hyperparameters are adjusted in k-1 of the dataset and tested on the excluded proportion. This is repeated until optimal performance is reached.
Feature selection	Occasionally including all variables or predictors can reduce predictive performance of a model. Using statistical techniques such as random forest, you can select which predictors are the most powerful. The other variables are excluded as they may add 'noise'.
Ensemble method and principal component analysis	Principal component analysis is a method to reduce the number of inputs into models by combining these variables into new variables, termed feature extraction. These new variables are ordered by how well they predict the dependent variable whilst dropping the least important.

Supplementary Table 3- Missing data and predictors included in each model

Missing data	Number missing	Percentage missing values of Dataset 1	Included in Dataset 1	Included in Dataset 2
Genetic platform	0	0%	Yes	Yes
tPB2	4820	60%		
tPNa	4853	60%		
tPNf	357	4%	Yes	Yes
t2	168	2%	Yes	Yes
t3	136	1.7%	Yes	Yes
t4	152	1.8%	Yes	Yes
t5	169	2.1%	Yes	Yes
t6	184	2.3%	Yes	Yes
t7	230	2.9%	Yes	Yes
t8	295	3.7%	Yes	Yes
t9	1274	15.9%		
tSC	397	4.9%	Yes	Yes
tM	341	4.2%	Yes	Yes
tSB	290	3.6%	Yes	Yes
tB	359	4.5%	Yes	Yes
tEB	6334	78.9%		
tHB	3433	42.5%		
Morula fragmentation	4936	61.5%		
Cleavage stage fragmentation	398	4.95%	Yes	Yes
Blastocyst grading (FB, EXB, HB)	4767	59.4%		Yes
ICM grade	4569	56.9%		Yes
Trophectoderm grade	4687	58.4%		Yes
Morula grade	4067	50.7%		Yes
IVF or ICSI	0	0%	Yes	Yes
Sperm concentration	246	3.1%	Yes	Yes
Sperm progressive motility	229	2.8%	Yes	Yes
Age of oocyte provider	251	3.1%	Yes	Yes
BMI	1870	23.3%		
AMH	4967	61.9%		
AFC	2774	34.6%		
Number of oocytes retrieved	0	0%	Yes	Yes
Sperm provider age	398	4.9%	Yes	Yes
Protocol type	0	0%	Yes	Yes
FSH dose	389	4.8%	Yes	Yes

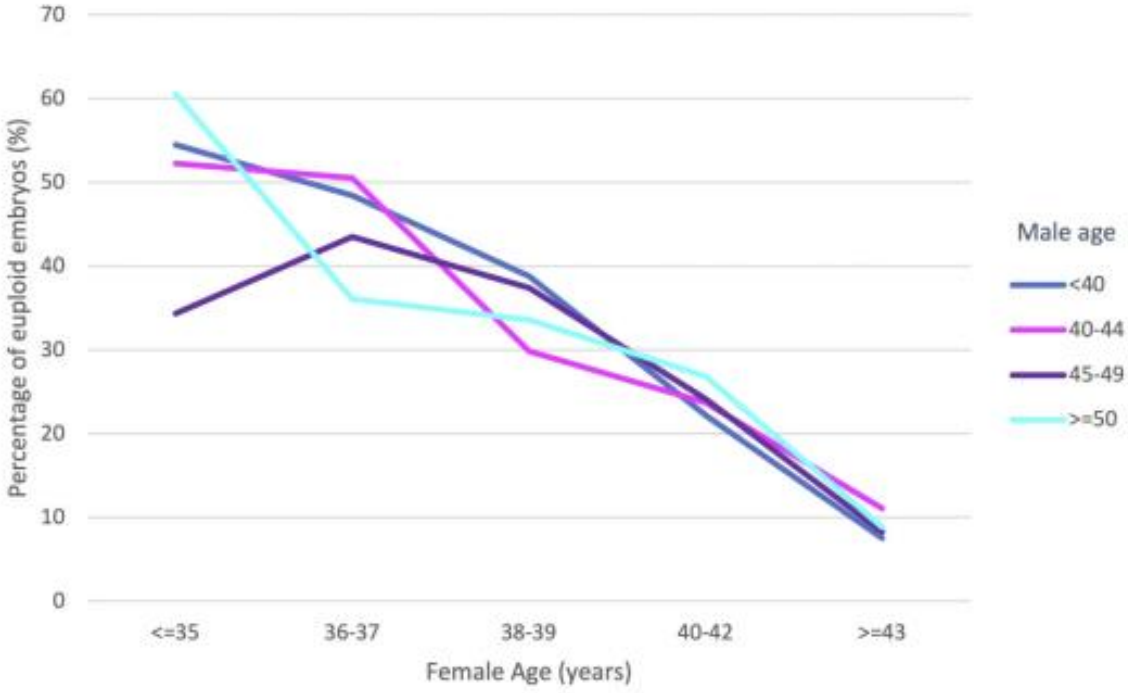
Supplementary Table 4- Full multivariable mixed effects logistic regression model prior to backward elimination (dataset I; target= euploid)

Covariate	Odds ratio	Lower 95% CI	Upper 95% CI	*P*-value
Genetic platform	1.061	0.744	1.507	0.750
tPNf	0.994	0.932	1.061	0.862
t2	1.022	0.962	1.084	0.485
t3	0.994	0.967	1.023	0.712
t4	0.992	0.963	1.021	0.567
t5	1.000	0.983	1.018	0.916
t6	1.017	0.995	1.039	0.123
t7	0.986	0.968	1.004	0.136
t8	0.997	0.986	1.008	0.641
tSC	1.009	0.998	1.021	0.102
tM	1.022	1.008	1.036	**0.002**
tSB	1.011	0.992	1.029	0.234
tB	0.945	0.931	0.959	**0.000**
ICSI or IVF	0.948	0.830	1.083	0.431
Sperm concentration	0.999	0.996	1.001	0.621
Sperm progressive motility	1.001	0.997	1.006	0.350
Embryo fragmentation	1.052	0.940	1.178	0.373
FSH dose	0.919	0.855	0.987	**0.022**
Oocyte provider age	0.839	0.821	0.858	**0.000**
Oocytes retrieved	0.995	0.986	1.004	0.318
Sperm provider age	1.007	0.993	1.021	0.293
Long or short protocol	0.825	0.692	0.983	**0.032**
Constant or intercept	3673.53	1107.58	12203.96	0.000

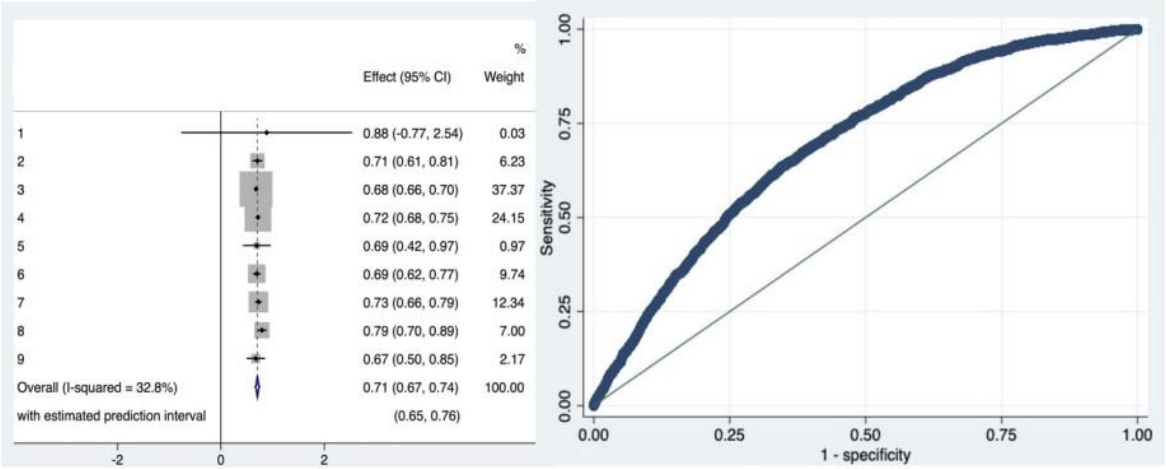
Supplementary Table 5- Final multivariable mixed effects logistic regression model after backward elimination (Dataset I; target=euploid)

Covariate	Odds ratio	Lower 95% CI	Upper 95% CI	*P*-value
t6	1.019	1.003	1.036	0.014
t7	0.982	0.969	0.995	0.008
tSC	1.001	0.998	1.02	0.085
tM	1.029	1.017	1.04	0.000
tB	0.948	0.939	0.957	0.000
FSH dose	0.928	0.869	0.991	0.026
Oocyte provider age	0.842	0.826	0.857	0.000
Long or short protocol	0.810	0.689	0.952	0.011
Constant or intercept	5114.5	1956.99	13366.71	0.000

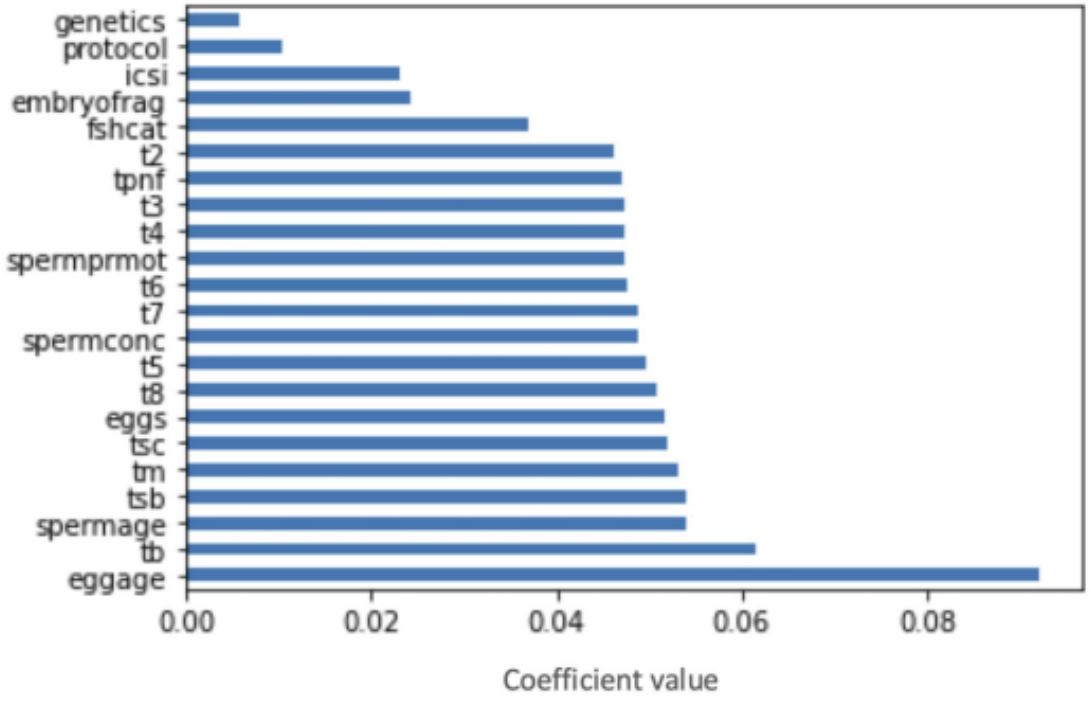
Supplementary Figure 1: Graph to show the relationship between male age and percentage of euploid embryos when stratified for female age



Supplementary Figure 2: Pooled AUC for internal-external cross-validation for Logistic Regression Model (Dataset I)



Supplementary Figure 3: Feature Importance for artificial intelligence modelling using Extra Trees Classifier



genetics: genetic platform used; protocol: long or short protocol; embryofrag: percentage of embryo cleavage stage fragmentation; fshcat: FSH dose; tn: time to n cells; tPNf: time to pronuclear fade; tSC: time to start of compaction; tM: time to formation of morula; tSB: time to start of blastulation; tB: time to formation of blastocyst; spermprmot: sperm progressive motility; spermconc: sperm concentration; spermage: sperm provider age; eggs: number of eggs retrieved.

APPENDIX III- SUPPLEMENTARY DATA FOR CHAPTER 4

Supplementary Table 1: Percentage of embryos with each risk score stratified by egg provider age

	PREFER						PREFER-MK only					
Egg provider age	Very High risk	High risk	Mod. High Risk	Mod. Low Risk	Low Risk	Very Low risk	Very High risk	High risk	Mod. High Risk	Mod. Low Risk	Low Risk	Very Low risk
<35	0.4%	5%	14%	25%	28%	28%	14%	16%	17%	17%	17%	18%
35-37	16%	44%	32%	8%	0.2%	0%	18%	17%	17%	17%	17%	16%
38-39	69%	29%	2%	0%	0%	0%	21%	20%	18%	14%	15%	12%
40-42	95%	5%	0%	0%	0%	0%	27%	18%	13%	15%	12%	15%
>=43	100%	0%	0%	0%	0%	0%	25%	22%	7%	18%	14%	14%

Supplementary Table 2: Sensitivity analyses adjusting for embryo quality

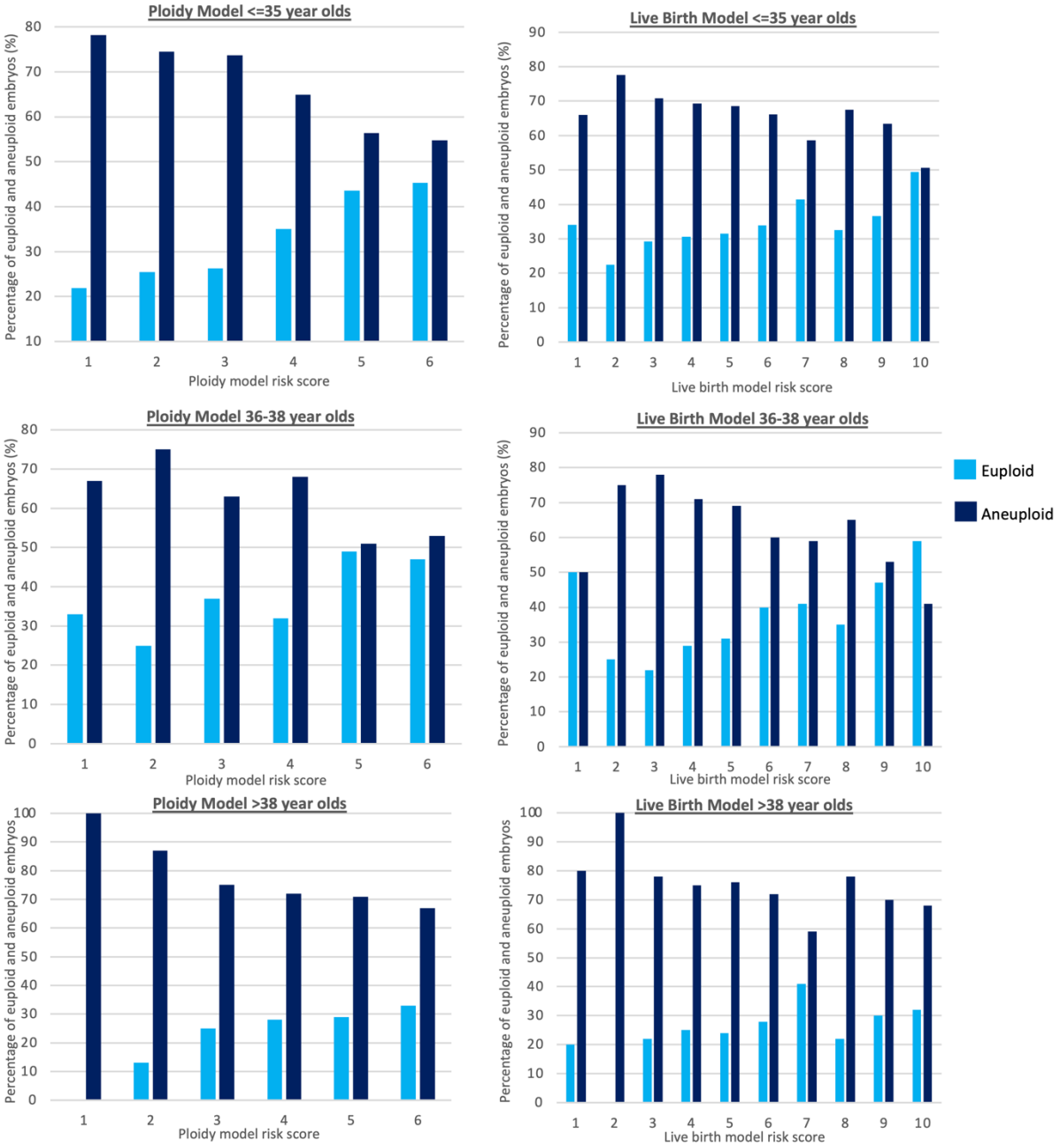
	PREFER		PREFER-MK (Morphokinetics only)	
Outcome	High risk aneuploid vs. Moderate risk (adjusted OR; 95% CI; p value)	High risk aneuploid vs. Low risk (adjusted OR; 95% CI; p value)	High risk aneuploid vs. Moderate risk (adjusted OR; 95% CI; p value	High risk aneuploid vs. Low risk (adjusted OR; 95% CI; p value)
Miscarriage rate	0.63; 95% CI 0.46-0.88; p<0.001	0.53; 95% CI 0.37-0.75; p=0.005	0.9; 95% CI; p=0.53	1.17; 95% CI 0.86-1.61; p=0.3
Biochemical pregnancy rate	1.65; 95% CI 1.38-1.97; p<0.001	1.8; 95% CI 1.48-2.2; p<0.001	1.35; 95% CI 1.14-1.6; p<0.001	1.49; 95% CI 1.14-1.6; p<0.001
Clinical pregnancy rate	1.64; 95% CI 1.37-1.95; p<0.001	1.87; 95% CI 1.54-2.27; p<0.001	1.48; 95% CI 1.26-1.76; p<0.001	1.58; 95% CI 1.33-1.89; p<0.001
Live birth rate	1.77; 95% CI 1.48-2.11; p<0.001	2.07; 95% CI 1.71-2.51; p<0.001	1.45; 95% CI 1.23-1.72; p<0.001	1.54; 95% CI 1.25-1.72; p<0.001

Supplementary Table 3: Sensitivity analyses comparing IVF vs. ICSI embryos for PREFER

IVF										
	PREFER					PREFER-MK (Morphokinetics only)				
Outcome	High risk	Moderate risk	Low risk	High risk vs. Mod. risk (adj OR; 95% CI; p value)	High risk vs. Low risk (adj OR; 95% CI; p value)	High risk	Moderate Risk	Low risk	High risk vs. Mod. risk (adj OR; 95% CI; p value)	High risk vs. Low risk (adj OR; 95% CI; p value)
Miscarriage rate	30/163 (18.4%)	44/250 (17.6%)	28/241 (11.6%)	0.97; 95% CI 0.57-1.65; p=0.91	0.57; 95% CI 0.3-1.07; p=0.81	24/153 (16%)	29/227 (13%)	49/274 (18%)	0.81; 95% CI 0.44-1.47; p=0.49	1.3; 95% CI 0.75-2.3; p=0.34
Biochemical pregnancy rate	190/385 (49%)	275/414 (66%)	264/380 (69%)	2.02; 95% CI 1.49-2.73; p<0.001	2.27; 95% CI 1.61-3.19; p<0.001	175/326 (54%)	252/418 (60%)	302/435 (69%)	1.3; 95% CI 0.94-1.73; p=0.1	1.9; 95% CI 1.4-2.6; p<0.001
Clinical pregnancy rate	163/385 (42%)	250/414 (60.4%)	241/380 (63%)	2.1; 95% CI 1.56-2.82; <0.001	2.41; 95% CI 1.72-3.4; p<0.001	153/326 (47%)	227/418 (54%)	274/435 (63%)	1.3; 95% CI 0.97-1.8; p=0.07	1.86; 95% CI 1.37-2.52; p<0.001
Live birth rate	132/385 (34%)	205/414 (49%)	211/380 (55%)	1.86; 95% CI 1.38-2.51; <0.001	2.43; 95% CI 1.74-3.4; p<0.001	128/326 (39%)	197/418 (47%)	223/435 (51%)	1.35; 95% CI 1.01-1.8; p=0.048	1.54; 95% CI 1.14-2.09; p<0.001
ICSI										
	PREFER					PREFER-MK (Morphokinetics only)				
Miscarriage rate	84/367 (22.9%)	56/450 (12.4%)	65/496 (13%)	0.49; 95% CI 0.31-0.72; p <0.001	0.51; 95% CI 0.36-0.78; p=0.002	66/398 (17%)	67/461 (15%)	72/456 (16%)	0.91; 95% CI 0.62-1.32; p=0.62	0.97; 95% CI 0.67-1.41; p=0.89
Biochemical pregnancy rate	417/810 (51.5%)	503/782 (64.3%)	546/816 (66.9%)	1.66; 95% CI 1.35-2.05; p<0.001	1.84; 95% CI 1.48-2.3; p<0.001	465/869 (54%)	500/778 (64%)	501/761 (66%)	1.49; 95% CI 1.229-1.83; p<0.001	1.57; 95% CI 1.27-1.92; p<0.001
Clinical pregnancy rate	367/810 (45.3%)	450/782 (57.5%)	496/816 (60.8%)	1.61; 95% CI 1.31-1.98; p<0.001	1.86; 95% CI 1.49-2.31; p<0.001	396/869 (46%)	461/778 (59%)	456/761 (60%)	1.68; 95% CI 1.38-2.06; p<0.001	1.69; 95% CI 1.38-2.08; p<0.001
Live birth rate	277/810 (34%)	391/782 (50%)	427/816 (52%)	1.89; 95% CI 1.53-2.32; p<0.001	2.1; 95% CI 1.68-2.605; p<0.001	327/869 (38%)	392/778 (50%)	422/761 (55%)	1.52; 95% CI 1.24-1.86; p<0.001	1.6; 95% CI 1.31-1.96; p<0.001

APPENDIX III- SUPPLEMENTARY DATA FOR CHAPTER 5

Supplementary Figure S1: Subgroup analysis of the association between ploidy and LB model risk scores and the proportion of euploid and aneuploid embryos



APPENDIX IV- REFERENCES

REFERENCES FOR CHAPTER 1

Alpha Scientists in Reproductive M and Embryology ESIGo. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting†. *Human Reproduction* 2011: **26**; 1270-1283.

Andriani GA, Almeida VP, Faggioli F, Mauro M, Tsai WL, Santambrogio L, Maslov A, Gadina M, Campisi J, Vijg J*, et al.* Whole Chromosome Instability induces senescence and promotes SASP. *Scientific Reports* 2016: **6**; 35218.

Bhatt S, Marchetto N, Roy J, Morelli S, and McGovern P. Pregnancy outcomes following in vitro fertilization frozen embryo transfer (IVF-FET) with or without preimplantation genetic testing for aneuploidy (PGT-A) in women with recurrent pregnancy loss (RPL): a SART-CORS study. *Human Reproduction* 2021: **36**; 2339-2344.

Cornelisse S, Zagers M, Kostova E, Fleischer K, van Wely M, Mastenbroek S. Preimplantation genetic testing for aneuploidies (abnormal number of chormosomes) in in vitro fertilisation. Cochrane Database of Systematic Reviews 2020, Issue 9. Art. No.:CD005281.

Coticchio G, Barrie A, Lagalla C, Borini A, Fishel S, Griffin D, and Campbell A. Plasticity of the human preimplantation embryo: developmental dogmas, variations on themes and self-correction. *Hum Reprod Update* 2021: **27**; 848-865.

Cummins J, Breen T, Harrison K, Shaw J, Wilson L, and Hennessey J. A formula for scoring human embryo growth rates in in vitro fertilization: its value in predicting pregnancy and in comparison with visual estimates of embryo quality. *Journal of In Vitro Fertilization and Embryo Transfer* 1986: **3**; 284-295.

Esteves SC, Carvalho JF, Martinhago CD, Melo AA, Bento FC, Humaidan P, and Alviggi C. Estimation of age-dependent decrease in blastocyst euploidy by next generation sequencing: development of a novel prediction model. *Panminerva Med* 2019: **61**; 3-10.

Ferrick L, Lee YSL, and Gardner DK. Metabolic activity of human blastocysts correlates with their morphokinetics, morphological grade, KIDScore and artificial intelligence ranking. *Human Reproduction* 2020: **35**; 2004-2016.

Gardner DK, Lane M, Stevens J, Schlenker T, and Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertil Steril* 2000: **73**; 1155-1158.

Griffin DK. Why PGT-A, most likely, improves IVF success. *Reproductive BioMedicine Online* 2022: **45**; 633-637.

Ho JR, Arrach N, Rhodes-Long K, Salem W, McGinnis LK, Chung K, Bendikson KA, Paulson RJ, and Ahmady A. Blastulation timing is associated with differential mitochondrial content in euploid embryos. *J Assist Reprod Genet* 2018: **35**; 711-720.
Mantikou E, Wong KM, Repping S, and Mastenbroek S. Molecular origin of mitotic aneuploidies in preimplantation embryos. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 2012: **1822**; 1921-1930.
Mio Y and Maeda K. Time-lapse cinematography of dynamic changes occurring during in vitro development of human embryos. *American journal of obstetrics and gynecology* 2008: **199**; 660. e661-660. e665.

Neal SA, Morin SJ, Franasiak JM, Goodman LR, Juneau CR, Forman EJ, Werner MD, and Scott RT. Preimplantation genetic testing for aneuploidy is cost-effective, shortens treatment time, and reduces the risk of failed embryo transfer and clinical miscarriage. *Fertility and Sterility* 2018: **110**; 896-904.

Sanders KD, Silvestri G, Gordon T, and Griffin DK. Analysis of IVF live birth outcomes with and without preimplantation genetic testing for aneuploidy (PGT-A): UK Human Fertilisation and Embryology Authority data collection 2016-2018. *J Assist Reprod Genet* 2021: **38**; 3277-3285.

Santaguida S, Richardson A, Iyer DR, M'Saad O, Zasadil L, Knouse KA, Wong YL, Rhind N, Desai A, and Amon A. Chromosome Mis-segregation Generates Cell-Cycle-Arrested Cells with Complex Karyotypes that Are Eliminated by the Immune System. *Dev Cell* 2017: **41**; 638-651.e635.

Seibel MM, Smith DM, Levesque L, Borten M, and Taymor ML. The temporal relationship between the luteinizing hormone surge and human oocyte maturation. *American journal of obstetrics and gynecology* 1982: **142**; 568-572.
Simopoulou M, Sfakianoudis K, Maziotis E, Tsioulou P, Grigoriadis S, Rapani A, Giannelou P, Asimakopoulou M, Kokkali G, and Pantou A. PGT-A: who and when? A systematic review and network meta-analysis of RCTs. *Journal of Assisted Reproduction and Genetics* 2021: **38**; 1939-1957.

Shapiro B, Melody A, Verma K, Raman A, Garner F, Aguirre M, Kayre L, Bedient C. The effect of ovarfian follicle size on oocyte and embryology outcomes. Fertility and Stwerility. Volume 117, Issue 6, June 2022; 1170-1176.

Tiegs AW, Tao X, Zhan Y, Whitehead C, Kim J, Hanson B, Osman E, Kim TJ, Patounakis G, Gutmann J*, et al.* A multicenter, prospective, blinded, nonselection study evaluating the predictive value of an aneuploid diagnosis using a targeted next-generation sequencing–based preimplantation genetic testing for aneuploidy assay and impact of biopsy. *Fertility and Sterility* 2021: **115**; 627-637.

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, and Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007: **370**; 1453-1457.

Williams BR, Prabhu VR, Hunter KE, Glazier CM, Whittaker CA, Housman DE, and Amon A. Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. *Science* 2008: **322**; 703-709.

Zhu J, Tsai H-J, Gordon MR, and Li R. Cellular stress associated with aneuploidy. *Developmental cell* 2018: **44**; 420-431.

REFERENCES FOR CHAPTER 5

Ahlström A, Lundin K, Lind AK, Gunnarsson K, Westlander G, Park H, Thurin-Kjellberg A, Thorsteinsdottir SA, Einarsson S, Åström M *et al.* A double-blind randomized controlled trial investigating a time-lapse algorithm for selecting Day 5 blastocysts for transfer. *Hum Reprod* 2022;37: 708-717.

Arce JC, Ziebe S, Lundin K, Janssens R, Helmgaard L, Sørensen P. Interobserver agreement and intraobserver reproducibility of embryo quality assessments. *Hum Reprod* 2006;21: 2141-2148.

Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database of Systematic Reviews* 2019.

Bamford T, Barrie A, Montgomery S, Dhillon-Smith R, Campbell A, Easter C, Coomarasamy A. Morphological and morphokinetic associations with aneuploidy: a systematic review and meta-analysis. *Human Reproduction Update* 2022.

Bamford T, Easter C, Montgomery S, Smith R, Dhillon-Smith RK, Barrie A, Campbell A, Coomarasamy A. A comparison of 12 machine learning models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos. *Human Reproduction* 2023.

Bamford T, Smith R, Easter C, Dhillon-Smith R, Barrie A, Montgomery S, Campbell A, Coomarasamy A. Association between a morphokinetic ploidy prediction model risk score and miscarriage and live birth: a multicentre cohort study. *Fertility and Sterility* 2023.

Barrie A. Design, implementation and results of a group-wide, embryo morphokinetic annotation quality assurance scheme across ten fertility clinics. Fertility 2021 Barriers and breakthroughs. 6–10th January 2021 Online. *Human Fertility* 2021;24: 46-69.

Barrie A, McDowell G, Troup S. An investigation into the effect of potential confounding patient and treatment parameters on human embryo morphokinetics. *Fertility and Sterility* 2021;115: 1014-1022.

Barrie A, Homburg R, McDowell G, Brown J, Kingsland C, Troup S. Examining the efficacy of six published time-lapse imaging embryo selection algorithms to predict

implantation to demonstrate the need for the development of specific, in-house morphokinetic selection algorithms. *Fertil Steril* 2017;107: 613-621.

Basile N, Nogales Mdel C, Bronet F, Florensa M, Riqueiros M, Rodrigo L, Garcia-Velasco J, Meseguer M. Increasing the probability of selecting chromosomally normal embryos by time-lapse morphokinetics analysis. *Fertil Steril* 2014;101: 699-704.

Baxter Bendus AE, Mayer JF, Shipley SK, Catherino WH. Interobserver and intraobserver variation in day 3 embryo grading. *Fertil Steril* 2006;86: 1608-1615.

Bosch E, Bori L, Valera MÁ, Delgado A, Naranjo V, Meseguer M. P-186 Combination of proteomics and automatic scoring using artificial neural networks to detect aneuploid embryos. *Human Reproduction* 2023;38.

Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Hickman CF. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. *Reprod Biomed Online* 2013;26: 477-485.

Campbell AJ, Petersen Dr BM, Smith R, Barrie A. Prediction of blastulation, embryo utilisation and live birth from single morphological or morphokinetic variables: analysis of 31,323 embryos gives insights for selection and algorithm development. Fertility and Sterility 2022;118:3138-138.

Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reprod Biomed Online* 2020;41: 585-593.

Chawla M, Fakih M, Shunnar A, Bayram A, Hellani A, Perumal V, Divakaran J, Budak E. Morphokinetic analysis of cleavage stage embryos and its relationship to aneuploidy in a retrospective time-lapse imaging study. *J Assist Reprod Genet* 2015;32: 69-75.
Ciray HN, Campbell A, Agerholm IE, Aguilar J, Chamayou S, Esbert M, Sayed S. Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Hum Reprod* 2014;29: 2650-2660.

Coticchio G, Barrie A, Lagalla C, Borini A, Fishel S, Griffin D, Campbell A. Plasticity of the human preimplantation embryo: developmental dogmas, variations on themes and self-correction. *Hum Reprod Update* 2021;27: 848-865.

Del Carmen Nogales M, Bronet F, Basile N, Martinez EM, Linan A, Rodrigo L, Meseguer M. Type of chromosome abnormality affects embryo morphology dynamics. *Fertil Steril* 2017;107: 229-235 e222.

Desai N. RP. Is there an increase in aneuploidy rate with delayed blastulation, multinucleation or cleavage anomalies? *Fertility and sterility* 2016;106: e145.

Desai N, Ploskonka S, Goodman LR, Austin C, Goldberg J, Falcone T. Analysis of embryo morphokinetics, multinucleation and cleavage anomalies using continuous time-lapse monitoring in blastocyst transfer cycles. *Reproductive Biology and Endocrinology* 2014;12: 54.

Fishel S, Campbell A, Montgomery S, Smith R, Nice L, Duffy S, Jenner L, Berrisford K, Kellam L, Smith R *et al.* Time-lapse imaging algorithms rank human preimplantation embryos according to the probability of live birth. *Reprod Biomed Online* 2018;37: 304-313.

Fréour T, Dessolle L, Lammers J, Lattes S, Barrière P. Comparison of embryo morphokinetics after in vitro fertilization-intracytoplasmic sperm injection in smoking and nonsmoking women. *Fertil Steril* 2013;99: 1944-1950.

Gardner D. K. SWB. “In vitro culture of human blastocyst,” in Towards Reproductive Certainty: Infertility and Genetics Beyond In Mortimer D. (Carnforth: Parthenon Press;) (ed). 1999.

Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertil Steril* 2000;73: 1155-1158.

Goodman LR, Goldberg J, Falcone T, Austin C, Desai N. Does the addition of time-lapse morphokinetics in the selection of embryos for transfer improve pregnancy rates? A randomized controlled trial. *Fertility and Sterility* 2016;105: 275-285.e210.

Griffin DK. Why PGT-A, most likely, improves IVF success. *Reproductive BioMedicine Online* 2022;45: 633-637.

Huang B, Zheng S, Ma B, Yang Y, Zhang S, Jin L. Using deep learning to predict the outcome of live birth from more than 10,000 embryo data. *BMC Pregnancy and Childbirth* 2022;22: 36.

Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can predict embryo ploidy status based on time-lapse data. *Reprod Biol Endocrinol* 2021;19: 185-185.

Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of IR techniques. *ACM transactions on information systems* 2002;20: 422-446.

Kaser DJ, Bormann CL, Missmer SA, Farland LV, Ginsburg ES, Racowsky C. A pilot randomized controlled trial of Day 3 single embryo transfer with adjunctive time-lapse selection versus Day 5 single embryo transfer with or without adjunctive time-lapse selection. *Hum Reprod* 2017;32: 1598-1603.

Kato K, Ueno S, Berntsen J, Ito M, Shimazaki K, Uchiyama K, Okimura T. Comparing prediction of ongoing pregnancy and live birth outcomes in patients with advanced and younger maternal age patients using KIDScore™ day 5: a large-cohort

retrospective study with single vitrified-warmed blastocyst transfer. *Reprod Biol Endocrinol* 2021;19: 98.

Kato K, Ueno S, Berntsen J, Kragh MF, Okimura T, Kuroda T. Does embryo categorization by existing artificial intelligence, morphokinetic or morphological embryo selection models correlate with blastocyst euploidy rates? *Reproductive BioMedicine Online* 2023;46: 274-281.

Katz-Jaffe MG, McReynolds S. Embryology in the era of proteomics. *Fertility and sterility* 2013;99: 1073-1077.

Krisher RL, Schoolcraft WB, Katz-Jaffe MG. Omics as a window to view embryo viability. *Fertility and Sterility* 2015;103: 333-341.

Lee CI, Chen CH, Huang CC, Cheng EH, Chen HH, Ho ST, Lin PY, Lee MS, Lee TH. Embryo morphokinetics is potentially associated with clinical outcomes of single-embryo transfers in preimplantation genetic testing for aneuploidy cycles. *Reprod Biomed Online* 2019;39: 569-579.

Meseguer M, Herrero J, Tejera A, Hilligsøe KM, Ramsing NB, Remohí J. The use of morphokinetics as a predictor of embryo implantation. *Hum Reprod* 2011;26: 2658-2671.

Mumusoglu S, Yarali I, Bozdag G, Ozdemir P, Polat M, Sokmensuer LK, Yarali H. Time-lapse morphokinetic assessment has low to moderate ability to predict euploidy when patient- and ovarian stimulation-related factors are taken into account with the use of clustered data analysis. *Fertil Steril* 2017;107: 413-421.e414.

Munné S, Kaplan B, Frattarelli JL, Child T, Nakhuda G, Shamma FN, Silverberg K, Kalista T, Handyside AH, Katz-Jaffe M *et al.* Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. *Fertil Steril* 2019;112: 1071-1079.e1077.

Paternot G, Wetzels AM, Thonon F, Vansteenbrugge A, Willemen D, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and interobserver analysis in the morphological assessment of early stage embryos during an IVF procedure: a multicentre study. *Reprod Biol Endocrinol* 2011;9: 127.

Quinn MM, Marsh P, Ribeiro S, Simbulan RK, Rosen MP. A deep dive into the morphokinetics and ploidy of low-quality blastocysts. *F&S Reports* 2022;3: 231-236.

Reignier A, Lammers J, Barriere P, Freour T. Can time-lapse parameters predict embryo ploidy? A systematic review. *Reprod Biomed Online* 2018;36: 380-387.

Rubio I, Galán A, Larreategui Z, Ayerdi F, Bellver J, Herrero J, Meseguer M. Clinical validation of embryo culture and selection by morphokinetic analysis: a randomized,

controlled trial of the EmbryoScope. *Fertility and sterility* 2014;102: 1287-1294. e1285.

Sanders KD, Silvestri G, Gordon T, Griffin DK. Analysis of IVF live birth outcomes with and without preimplantation genetic testing for aneuploidy (PGT-A): UK Human Fertilisation and Embryology Authority data collection 2016–2018. *Journal of Assisted Reproduction and Genetics* 2021;38: 3277-3285.

Storr A, Venetis C, Cooke S, Kilani S, Ledger W. Time-lapse algorithms and morphological selection of day-5 embryos for transfer: a preclinical validation study. *Fertil Steril* 2018;109: 276-283.e273.

Theobald R, SenGupta S, Harper J. The status of preimplantation genetic testing in the UK and USA. *Hum Reprod* 2020;35: 986-998.

Tiegs AW, Tao X, Zhan Y, Whitehead C, Kim J, Hanson B, Osman E, Kim TJ, Patounakis G, Gutmann J *et al.* A multicenter, prospective, blinded, nonselection study evaluating the predictive value of an aneuploid diagnosis using a targeted next-generation sequencing–based preimplantation genetic testing for aneuploidy assay and impact of biopsy. *Fertility and Sterility* 2021;115: 627-637.

Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Human Reproduction* 2019;34: 1011-1018.

Uyar A, Bener A, Ciray HN. Predictive Modeling of Implantation Outcome in an In Vitro Fertilization Setting: An Application of Machine Learning Methods. *Med Decis Making* 2015;35: 714-725.

Verpoest W, Staessen C, Bossuyt PM, Goossens V, Altarescu G, Bonduelle M, Devesa M, Eldar-Geva T, Gianaroli L, Griesinger G *et al.* Preimplantation genetic testing for aneuploidy by microarray analysis of polar bodies in advanced maternal age: a randomized clinical trial. *Human Reproduction* 2018;33: 1767-1776.

Yan J, Qin Y, Zhao H, Sun Y, Gong F, Li R, Sun X, Ling X, Li H, Hao C *et al.* Live Birth with or without Preimplantation Genetic Testing for Aneuploidy. *New England Journal of Medicine* 2021;385: 2047-2058.