ANALYSIS OF EMBRYO SCORING AND COMPARISON OF CLINIC PERFORMANCE IN IN VITRO FERTILIZATION

A Thesis

Presented to the Faculty of

California Polytechnic State University

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Industrial Engineering

by

James W. Whistler

March 2015

COMMITTEE MEMBERSHIP

TITLE: Analysis of Embryo Scoring and Comparison of Clinic Performance in In Vitro Fertilization

AUTHOR: James W. Whistler

DATE SUBMITTED: March 2015

COMMITTEE CHAIR: Dr. Jianbiao Pan, Professor of Industrial and Manufacturing Engineering

COMMITTEE MEMBER: Dr. Alex Steinleitner, Reproductive Endocrinologist

COMMITTEE MEMBER: Dr. Reza Pouraghabagher, Professor of Industrial and Manufacturing Engineering

ABSTRACT

Analysis of Embryo Scoring and Comparison of Clinic Performance in In Vitro Fertilization

by

James Whistler

Clinical Assisted Reproductive Technology (ART) practices seek to make improvements in embryo quality and resultant procedural success rates. There is a significant variance in live birth rates among clinics nationwide. The goal of this thesis is make comparisons of embryo quality among clinics and understand these differences. This analysis focuses on the stage between egg retrieval and embryo transfer. Because the currently accepted embryo scoring methods are not directly proportional to performance, a new scoring methodology is proposed and applied. Data provided by the Society for Assisted Reproductive Technology (SART) consisting of 36,836 patient cycles from 40 anonymous clinics nationwide is considered. After necessary reductions are made, the data is anatomized to link each embryo transferred to an implantation probability. A score is generated for each morphology grouping based on the average implantation rate of that group. This score is used as the basis for clinic comparisons. Top-performing clinics (in terms of live birth rates in patients aged <35 years old) are then shown to both produce embryos of higher score and achieve better results from embryos of identical morphology.

Keywords: IVF, Embryo Scoring, Clinic Comparisons, Contingency Tables, Data Mining

ACKNOWLEDGMENTS

Working on this project was a great experience for me. I am very grateful for Dr. Pan reaching out to me about the opportunity to work on this project. It certainly proved to be the challenge I was looking for and more. It was a pleasure working with both Dr. Pan and Dr. Steinleitner throughout the process. I really enjoyed our spirited dialogues that frequently occurred in weekly meetings. I learned a tremendous amount, not just about statistics and In Vitro Fertilization (IVF), but also about working in teams, presenting information and professionalism. I thank Dr. Pouraghabagher for sharing his wisdom and sitting on the committee. I would like to also thank Sofia Soto and Heidi Richardson for their work on their senior project that helped contribute to this report. Lastly, I would like to thank my parents and family for their unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Every year, many people turn to Assistive Reproductive Technology (ART) to seek treatment for fertility problems. A method that has become increasingly viable is In-vitro Fertilization (IVF). The IVF process is essentially:

1. The woman is given a drug treatment to hyper-stimulate ovarian release of eggs
2. These eggs are fertilized in petri dishes and grown for 3 to 5 days into Embryos
3. Some of these Embryos are transferred to the uterus
4. If the transfer is successful, the remainder of the pregnancy continues as would a conventional pregnancy

These steps are illustrated in Figure 1 below:

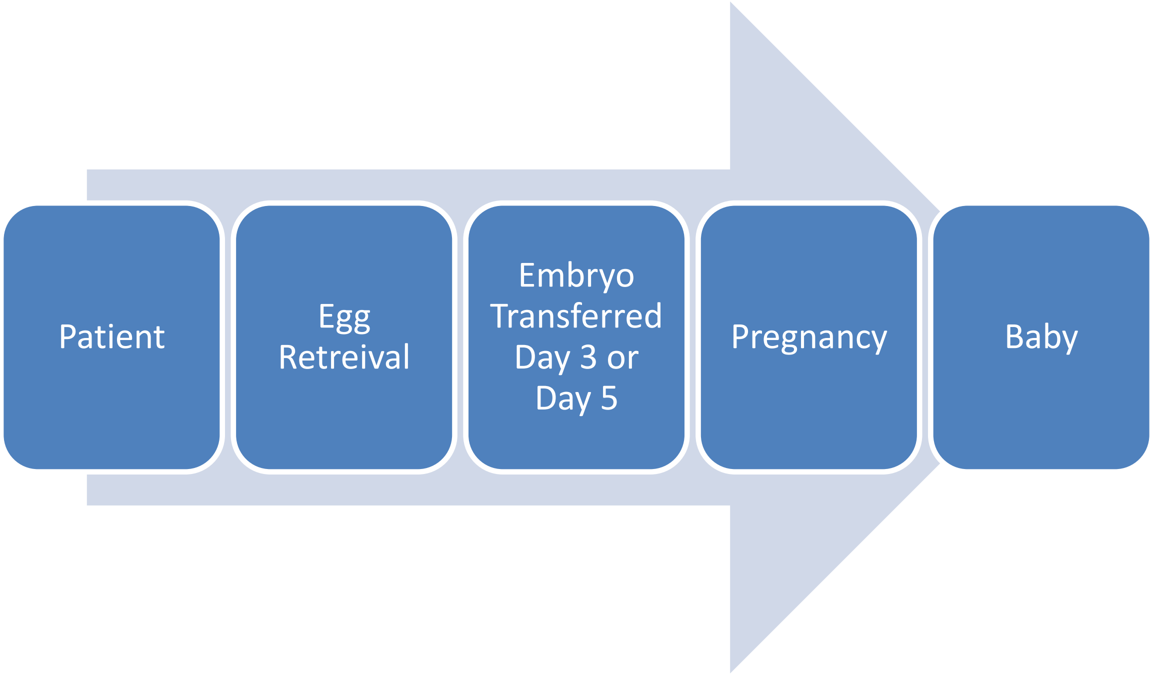


Figure 1: Process Steps

This procedure is recognized and performed globally. There are currently 440 registered clinics in the US (“SART National Summary”, 2015). Each of these clinics aims to maximize positive patient outcomes, namely the delivery of healthy babies. At a high level the process is very similar at each clinic; however, there is a huge disparity in live birth rates among clinics. The top performing clinics reporting live birth rates for patients aged less than 35 in over 70% of cycles, compared to the lowest performing clinics reporting less than 30% (“Fertility Success Rates”, 2013).

This thesis aims to take a new approach in analyzing the differences between clinics by considering ART practice as a manufacturing process wherein inputs of "materials" (patients, gametes, reagents, consumables) and labor are employed under specific environmental conditions to produce high quality embryos to initiate pregnancy. To facilitate such a comparison and provide a framework for future quality monitoring and improvement, a new metric of embryo quality is developed. This research was conducted through the exploration of a large dataset provided by the Society for Assisted Reproductive Technology (SART).

This thesis employs data mining and statistical analysis techniques to show correlations and trends in the data as well as differences between groups. The power and significance of the conclusions are made possible by the large quantity of cases that are available for examination.

The IVF process can be modeled as a manufacturing process: typical manufacturing processes aim to reduce defects and improve overall quality. In the context of IVF, this means maximizing the quality of Embryos produced. The first step in moving toward quality improvements lies in understanding variation in the process. Later designed experiments can be used to attach this variation to causation. Subsequently, having consistently higher success rates adds additional value to the customers: in this case, the patients. This allows clinics who can understand the causes for variation and improve upon them a competitive advantage. This competition will drive improvements across all clinics.

The particular stage of interest of this thesis is that between egg retrieval and embryo transfer. In this stage, practitioners want to determine the potential of embryos based on what they look like under a microscope. From looking at embryos closely, values of morphological categories can be measured objectively, such as the symmetry or fragmentation of the embryo. The challenge is translating an array of values into one score that reflects the potential of that embryo to result in a successful pregnancy. Several scoring methods have been shown to correlate with performance; however, none are directly derived from results. These are detailed in the literature review. This exploratory study aims to utilize a large dataset to develop a new 1-dimensional Embryo Score based upon actual performance.

After a reliable scoring method has been established, the large dataset can be used to compare performance among clinics. Of course, confounding patient factors must be considered. After significant factors are removed, how the mean embryo score differs by clinic is examined. This measures the quality of embryos produced by clinics. Additionally, the difference among clinics in performance of embryos with identical morphology (and thus the same score) is explored.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This literature review serves multiple purposes and accordingly, is divided into two main sections: IVF process understanding and exploration of existing research. A general overview is provided before diving into more detailed considerations. The IVF process understanding begins with a high level exploration of how Assisted Reproductive Technology has evolved over time and then focuses on that have changed over the years. It discusses the evolution of Assisted Reproductive Technology and the differences between the early days of IVF and today's practices.

Prominent research objectives in industry studies and reports are reviewed. Also, critical methodologies and results are presented, making evident some gaps in existing literature. The literature review concludes by discussing how this thesis targets some of the existing gaps.

2.2 IVF Process

In 2011, over 150,000 patients underwent IVF treatment in the United States ("Assisted Reproductive Technology", 2013). It is estimated that over 5 million babies have been born worldwide via IVF since Louise Brown, the first test-tube baby, was born in 1978 ("ART fact sheet", 2014). Infertility refers to a biological inability to conceive after regular, unprotected sex, or a female who can't carry a pregnancy to full term (Nordqvist, 2014). Ovulatory disorders are one of the most common causes for infertility in women, accounting for 30% of women's infertility. Ovulation is the monthly release of an egg. In some cases, the woman never releases eggs, and in others, the woman does not release eggs during some cycles (Nordqvist, 2014). Ovulatory disorders are caused by issues including hormonal problems, damaged or scarred ovaries, premature menopause, and follicle problems (“What Causes Female Infertility”, 2015). Ovulatory problems can result in damaged ovaries, follicles, the production of immature eggs, or no eggs. In these cases, chances of fertilization become nearly nonexistent (“What Causes Female Infertility”, 2015).

Another factor affecting about 25% of infertile couples is poorly functioning fallopian tubes, or tubal disease. It is caused by infections, abdominal diseases, previous surgeries, ectopic pregnancy, and congenital defects. The fallopian tubes can experience a range of issues from mild adhesions to complete blockage, in which case eggs can’t, or have trouble being released.

Endometriosis affects about 10% of all infertile couples. For women with endometriosis, their monthly chance of getting pregnant is reduced to 36%. (“What Causes Female Infertility”, 2015).

Other factors include obstruction of the uterus and fallopian tubes, congenital abnormalities, cervical mucus problems, and behavioral factors such as diet, exercise, smoking, and drugs, and environmental and occupational factors.

IVF is a multistep process, or cycle, in which eggs are extracted and fertilized with sperm in a lab. Once embryos develop, some are implanted in the woman’s uterus (typically 1-3) and if further viable embryos remain, they are stored (Christiano, 2015). This process can be used to treat some of the aforementioned infertility problems, including: blocked or damaged fallopian tubes, women with ovulation disorders, women who have had their fallopian tubes removed, women with severe endometriosis, genetic disorders, and unexplained infertility (Romito, 2013). The success of IVF treatment varies by age – yielding 41% pregnancy rate for women under age 35, 32% for women ages 35 to 37, and 23% for those ages 38 to 40 (Christiano, 2015). If successful, couples with serious fertility problems can become parents. However, treatments can be costly – averaging about $8,000 per cycle before medication – and can be physically demanding, requiring a regiment of fertility drugs before the start of each cycle (Christiano, 2015). These drug treatments can have serious implications for the mother and baby.

2.3 Existing Research

Challenges in the Field Today: Risks of Higher Order Multiple Pregnancies

Higher success rates in both oocyte and embryo production have led to increasingly alarming rates of twinning and Higher Order Multiple (HOM) pregnancies. Across the board, initiatives have been made to suggest moving in the direction of Single Embryo Transfer, thus reducing the chance of fraternal twinning to zero.

A consideration for the biological efficiency of IVF is not a new idea in general. Doctors and Researchers have expressed concern for the impacts of the drug therapy typically undergone by IVF patients. As mentioned in the IVF Process section, patients undergo drug therapy to hyperstimulate the ovaries, resulting in greater egg production. However, some claims have been made as to the "unnaturalness" of this drug therapy — in some cases releasing 30 or more eggs in one cycle as opposed to the average of 1, or rarely more than 3 in a natural ovulation period.

One study was performed to explore the high biological inefficiency of wasted oocytes (Patrizio & Sakkas, 2009). The researchers used data provided by SART to analyze efficiency of the process at each stage, most notably investigating the efficiency of each oocyte resulting in embryos and subsequently, successful live births.

Wasting a few eggs cells may be inefficient, does not pose a high biological risk. The larger concern here is the increased risk of twinning and particularly, HOM pregnancies. The rates of twinning and HOM pregnancies are significantly higher in the IVF process compared to traditional conception methodologies. With each additional baby delivered, there is heightened morbidity and mortality risk for both the mother and baby (Pector, 2005). Not to mention, the additional financial and emotional burden such cases may put on families, who are already spending a good deal of money on the procedure itself.

Due to the risks and costs associated with HOM pregnancies, initiatives have been made, across the board, to suggest moving in the direction of Single Embryo Transfer in order to reduce the chance of fraternal twinning to zero.

Transfer of a single embryo requires a high level of confidence in the ability of an embryo to implant. A study by Jungheim et al. (2010) discussed the benefits of shifting towards fewer and even single embryo transfer. They performed a survey to determine how well the American Society for Reproductive Medicine (ASRM) embryo transfer guidelines were being followed by practitioners nationwide, showing that 55% of responders will deviate from the guidelines upon patient request. This shouldn’t be surprising, as this procedure is typically not covered by insurance; patients making this investment want to have the best chance of success. They cite the extreme case of the Octomom, a well known case of a woman who had 6 embryos transferred, all of which implanted and two became identical twins. This was a statistical anomaly and seems only to be referenced in this paper to invoke an emotional response from the reader. Further research should be done to assess the probability and risks of HOM pregnancies in order to adequately define the upper and lower limits for embryo transfer.

A study examining the efficiency of oocytes becoming embryos by Patrizio and Sakkas notes that the vast majority of embryo transfers in the IVF process happen either 3 or 5 days after fertilization. An interesting finding within this dataset was that the percentage of day 3 transfers is higher as patients get older. This increase is from 68% in patients under 35 to 86.7% in patients over 42. This leads to a potential confounding factor when comparing Embryo quality across day 3 and day 5 transfers. (Patrizio & Sakkas, 2007)

Patrizio's study examined 572 retrieval cycles and found that 2252 oocytes resulted in usable embryos, which accounts for around 31% of total oocytes retrieved. From these, a final live birth rate of 6.8% per oocyte was reported. In patients over 40, birth rates per oocyte were below 1%. The authors seem to be appalled by this low efficiency rate. The ethical decision becomes whether or not wasting a dozen eggs to create a healthy baby that would not have been possible otherwise is certainly worth it. They do bring up a plausible hypothesis: some of the oocytes retrieved may be intrinsically abnormal and ovarian hyperstimulation may interfere with the intracellular events happening here. However, very little is known about the molecular dynamics taking place. This is an area that needs more exploration from a biological perspective.

Patrizio recommends minimal or mild ovarian stimulation, and further exploration of such practice. Perhaps his most important point is that producing a higher number of oocytes doesn't result in more viable embryos. To make this more powerful and build on his research, a designed experiment comparing low and high ovarian stimulation in an otherwise equivalent environment would be useful. The problem with just an exploratory study comparing results based on number of eggs produced could be confounded with any number of biological factors.

Patrizio's goal is to limit excessive oocyte production and embryo transfer. The practical challenge is the high degree of inconsistency in quality, making this a challenge without reducing live birth rates. He concludes by calling for a more standardized metric to measure IVF efficiency in hopes of promoting greater awareness and encouraging improvement.

Studies focusing on day 5 transfers of blastocysts indicate that transferring more than 2 embryos does not provide additional value (Stern et al., 2008). In patients where 3 and 4 embryos were transferred, lower pregnancy rates results. However, it is likely that these situations were a result of the morphologically poor embryos that caused the doctor to transfer so many (A. Steinleitner, personal communication, 2015) This goes along with the aforementioned idea that the presence of one strong embryo is more important than mean embryo score.

Many researchers have proposed that Single Embryo Transfer (SET) results in lower pregnancy rates with an odds ratio of .53 when compared to double embryo transfer (Bhattacharaya & Templeton, 2004). It is important to note that the risk of multiple pregnancy is decreased by a factor of 10 in randomized controlled trials (Bhattacharaya & Templeton, 2004). In order for SET to be more widely considered, probability of patient outcomes must be better understood.

Factors Affecting IVF Success

It is important to understand which factors influence the success of an IVF procedure. This paper will examine quality in clinics, so it is important to understand what variability exists in the inputs (the patients) and how they influence the procedure. It is also relevant to understand the ways in which patient related factors can be practically tracked. This allows the clinic to separate what factors they can and cannot affect.

One of the most widely recognized factors is decreased fertility as women age. A retrospective study of 878 IVF cycles indicated that fertility may decrease even sooner than the previously widely accepted mid 30's threshold. (Ziebe et al., 1997) This analysis studied 3 clinics over a four-year time span from 1993 to 1996. The metrics that researchers focused primarily on were the number of oocytes produced and number of cleavage state embryos produced in day 2. The researchers found a highly significant decrease in oocyte production as age increased. They also note that the stimulation dose is typically increased in older patients. This would suggest that the ovarian response is diminished even more than indicated by these results.

Additionally, they show significance in the decrease of the proportion of oocytes that are aspirated into cleavage stage embryos. In each of these cases, the significance is easy to demonstrate due to a large sample size. This is certainly valid when making comparisons of mean performance by age group. However, the linear model is not effective in predicting the performance (Number of oocytes produced and Cleavaged/Aspirated Ratio based on age). It is not listed in the study, but from the graphs included, it is evident that the R-squared value of the model is very low; thus the predictive utility is limited.

Further, the researchers zoom in on specific groups. For example, they show that there is a negative trend in the implantation rate of good quality 4-cell embryos as age increases. (Ziebe et al., 1997) This demonstrates that, all other factors considered equal, an increase in patient age will have a negative effect on embryo performance.

Another study was done to examine the effect of max Folicule-Stimulating Hormone (FSH) level on ART treatment outcomes. FSH helps control production of eggs by the ovaries. Considering a large sample from SART data of 19,682 cycles in 1999, it was shown that a significant negative correlation exists between maxFSH and treatment outcomes. (Frazier, L. , Grainger, D., Schieve, L., & Toner, J., 1999) In other words, as FSH level increases, pregnancy rate and live birth rate decrease. It will be interesting to compare these results with those in the 2009-2011 data examined in this paper.

Additionally the study found a weaker but still significant correlation between estradiol-17beta, commonly referred to as E_2 level and treatment outcomes. These correlations were shown individually in each age grouping. A noted limitation of this dataset is that the patients' max FSH level is tracked in each case, but the day of the cycle in which this information was obtained is not available. Therefore, it is unclear whether the data comes from a basal measurement on the second or third day of the cycle or from a Clomid Challenge Test (CCCT) result that is typically measured on day 10.

In contrast to Frazier et al.'s research, a 2003 study found that FSH level alone was not a significant factor in affecting treatment outcomes. However, FSH level was significant when considered jointly with age.

An excellent point is made in that patients will heavily weigh their chance of successful pregnancy when considering whether or not to undergo the IVF process. This information is critical for counseling patients. The more accurately practitioners can predict the expected success at each phase in the process, the more informed patients will be as a result.

One of the major factors impacting a patient's chance at success is their age. Ovarian reserve, defined as the quality and quantity of the remaining follicle pool (Chuang et al., 2003), is shown to diminish in most women during their mid to late 30s. Ovarian responsiveness to stimulating drugs has been shown to be a strong indicator of IVF treatment outcomes and is a useful metric to track.

Chuang's study (2003) focuses on the patients undergoing treatment at the National Taiwan University over a 5-year period spanning 1,045 treatment cycles. All cycles examined were the first undergone for each woman. This study split up patients into 3 age groupings (<35,35-39,>=40) and 2 basal FSH groupings (<10 mIU/mL and >10mlIU/mL) resulting in 6 combinations.

Based on a logistic regression analysis, they found age to be an independent predictor of pregnancy rate, but not basal FSH. When both are in the model together, the AUC (area under receiver operating characteristic curve) raises from .617 to .627 — a minimal improvement. However, age and basal FSH together make for the most effective predictor of number of oocytes collected. (Chuang et al., 2003) This study is demonstrative of the benefits of tracking data reliably over a long period of time in drawing significant conclusions. Chuang concludes by saying that although FSH is related to age, and each is related to ART success, the predictors are not perfectly correlated with outcomes, and more research is required.

Dr. Stern and her colleagues focus on patients over 38 years old in their aim to determine the optimal number of embryos to transfer in these situations examining factors such as prior pregnancy, FSH levels, number of oocytes retrieved, and number of embryos cryopreserved. They focused on these factors because the uncertainty is greater in older patients and the risks associated with HOM pregnancies are greater. Their goal was to develop an algorithm indicating the optimal number of embryos to transfer in each case. However, the study primarily looks at correlation between variables which are procedural inputs. An examination of each variable's effects on output would be more useful. Potential measures of output could be implantation or live birth rate.

Nonetheless, Stern et al. (2009) did find an interesting correlation between number of embryos cryopreserved and delivery rate. From a logistic regression analysis, they found a relative risk of pregnancy to be .77 in patients with less than 5 embryos compared to those with greater than 15 frozen with significance ($P<0.001$). This is certainly something that should be considered in future studies. Number of embryos cryopreserved seems to be a good indicator it is a good indicator of process stability. They also noted that some centers are "day 3 only", meaning they either don't have the capability or choose not to grow embryos beyond this. This is interesting information that could not be derived from the SART dataset, as clinics are not uniquely identified.

When they did look at output, this study used pregnancy rate by patient as a metric. This doesn't account for the number of embryos transferred and perhaps oversimplifies what would be better measured using implantation rate. Mentioned is the need to explore the impact of individual clinics due to different procedures and success rates.

Another group of factors that can be highly indicative of the IVF treatment outcomes are embryo morphological characteristics. Scoring based on these characteristics together is a focus of many studies. In particular, the effect of fragmentation in day 3 embryo success has been explored extensively. From a sample of 5,916 embryos, the average fragmentation was 15.4% on day 3, and 8.6% in those embryos that replaced. The mean implantation rate (IR) for this data set was 29.9% (Alikani et al., 1999). There is a high degree of significance in difference of embryo IR and pregnancy rate by fragmentation ($p<0.05$). However, Alikani points out that this is largely due to the extremely poor performance of embryos in the worst grouping, with greater than 35% fragmentation. The researchers found no correlation between fragmentation pattern and maternal age. They also found minimal impact of fragmentation removal, when it was possible. It is proposed that further research could be conducted on the impact of different patterns of the fragmentation rather than merely a 1-dimensional percentage.

How To Measure IVF Clinic Success

Dr. Gibbons and his colleagues (2007) bring a fantastic perspective to the major dilemma facing ART practitioners. They recognize the major goals that exist:

1. Improve Live Birth Rates from IVF process
2. Reduce the occurrence of twinning and HOM pregnancies

Goal 1 can be easily achieved by transferring more embryos and goal 2 can be achieved by transferring fewer. The challenge is to make simultaneous progress on both fronts. Thus, it is necessary to either improve the consistency of the embryos produced or develop a better way to predict the success of embryos. The first option requires biological or procedural advances while the second, merely a better understanding of the data that exists and perhaps additional information-tracking in the future. Better understanding and predicting the success clear target area for improvement.

Dr. Gibbons examined data from the ART nationally reported data and makes an astute observation: the lower performing clinics compensate for lower implantation rates by transferring more embryos. This results in higher twinning and triplet rates. He raises the question: “Should we be reporting implantation rates instead of live birth rates to more accurately reflect program quality?” (Gibbons et al., 2007). The counter consideration here is: does this merely reflect the patient quality of that clinic?

Another thing that Dr. Gibbons considers is the idea that patients are extremely sensitive to slight changes to implantation rate and, for example, will drive to the next city to use a clinic that has a 43% success rate vs. a 38%. This puts tremendous competitive pressure on clinics to produce a high birth rate and is likely what motivates the higher embryo transfer numbers in the poor clinics. He suggests instead showing number of standard deviations away from the national mean in order to downplay small differences. Dr. Gibbons is strongly against government regulation as the field is quickly changing and he feels that it does not provide adequate flexibility for individual situations and could therefore inflict unnecessary financial hardships on infertile couples.

In light of the data that Dr. Gibbons studied, it is clear that ART practitioners are motivated by self interest to improve their relative success. Accordingly, the measure of success that is most widely used should be consistent with making advancements toward the two goals outlined by Gibbons above with the appropriate weighting. In 2004, an important debate was conducted to achieve this, headed by Min et al. (2004).

ART is becoming better established and is no longer an experimental procedure as it was in the early 90s. In the early stages, the goal was to try to create a miracle in the form of a successful pregnancy for a previously infertile couple. Since then, success rates have increased markedly as techniques have been refined and improved. In 2004, a very important debate occurred in the community to figure out what the most relevant standard of success in assisted reproduction.

Min et al. (2004) argued that the singleton live birth rate should be used as the most relevant metric of ART success. In particular, they point to the statistic BESST (Birth Emphasizing a Successful Singleton at Term). The authors point out that there is a significantly higher risk of complications. Also, the cost of delivering HOM pregnancies is notably higher, at $170,282 for triplets, and $281,698 for quadruplets compared to $58,865 for twins (ESHRE, 2000). It is important to emphasize that they are considering a case of multiple live births to be the same value as an unsuccessful pregnancy or no pregnancy at all based on this metric.

Countering Min and his colleagues, Davies et al. (2004) contested that looking at the BESST measure alone biases the data because it is strongly influenced by extremes of age and embryos transferred, namely, patients over 35 or those who electively selected to transfer fewer embryos. It is not appropriate to measure the quality of the ART facility based on these factors. Additionally, the denominator comes from the number of cycles initiated. Since many cycles are unsuccessful in their early phases as they are unable to achieve a reasonable fertilization rate due to patient related factors.

Davies (2004) emphasizes that the goal is to measure and minimize the degree of burden or unwanted consequences and points out that it's difficult to measure this with only one parameter. For example, if implantation rates increase for a particular clinic, the overall pregnancy rate will increase, but so will the occurrence of multiple pregnancies — thus the BESST rate will change unpredictably.

Contributing to the discussion, Heijen et al. (2004) agreed that singleton live births are the goal, but the patient discomfort, risk of complications and costs should be considered. The measure of success should reward a clinic that produces a similar birth rate with milder ovarian stimulation and single embyro transfer compared to a clinic that transfers 3 or more embryos in young patients. A key point is that there is a dropout rate of approximately 25% after an unsuccessful process. The authors affirm that this is not only due to the cost factor or a poor prognosis, but also that the treatment process itself can be emotionally and physically stressful. In order to encourage clinicians to reduce the stressfulness of the process, the authors suggest that he denominator of success rate measure be number of treatments started rather than number of cycles.

Heijen (2004) mentions the idea of giving a higher value to the outcome of twins over an unsuccessful cycle rather than only crediting a singleton birth. For example, a singleton could be scored 1, and twins 0.5.

It is important to consider the implications of adopting a universal metric of IVF success. Clinics are businesses that compete with one another and will do what they can to improve this metric and increase their market share. Patients are motivated by clinics that they feel will give them the best shot at being successful in having a baby, so it's likely that they will likely pay more or travel a considerable distance. As a result, clinicians will certainly aim to improve their score.

Changing the metric from singleton births per cycle to singleton births per stared treatment helps encourage practitioners to shift towards elective single embryo transfer and milder ovarian stimulation. This advances the goals of decreasing negative consequences of IVF procedures such as HOM pregnancies or excessive stress on the mother.

Attempts at Developing an Effective Embryo Scoring System

It is useful to understand the relative value and the absolute value of an embryo based on its morphological characteristics. That being said, there has been significant research done in attempt to develop a method of scoring embryos based on easily-measurable data. However, to this day, there is no single widely-adopted method. Surprisingly, the study that has been most predominant in the community — and still relevant today — is from 1986. Given that IVF processes have evolved so much means that there are holes in what is established and new research is needed to fill these gaps. Reviewing some of the literature documenting attempts to develop a scoring method will show where the research stands.

A study done by Cummins et al. (1986) was one of the first attempts at establishing a reliable method with which to score embryos. This paper is extremely important in this literature review as it was the first and most widespread attempt at generating an embryo score and was widely accepted by many in the ART community. Their goal was to establish a general method for determining the quality of embryos based on visually apparent criteria, thus making predictions as to the performance of said embryos. They use two main criteria in scoring embryos: Embryo Development Rating (EDR), and a 1-4 scale of Embryo Quality (EQ) with 4 being the highest. The EDR is based on a regression model for the mean growth rate, and the visual quality (EQ) metric is derived from fragmentation, symmetry, and cytoplasm quality.

Cummins and his colleagues established groupings for EDR and EQ combinations and calculated “success rates” for each grouping. Success rate is not identical to implantation rate and is instead the probability that an embryo with this EQ/EDR combination will be associated with a successful pregnancy. This causes an inherent bias in their results and must have only been used due to an inability to track which embryo was successful in the situation of multiple embryo transfer. Considering this, the researchers looked at the 357 cases they had available with single embryo transfer (of which 33 were successful) and were able to show significance in some groups however the power of these results was limited due to the small sample size.

Still, on the basis of the success rate metric, comparisons can be made between EQ/EDR combination groups. Significant differences were found between many of the groupings. The highest rate of success in the category with EDR slightly above average and EQ is maximized (4). Also, the fastest growing embryos (EDR > 130) did not show an increased pregnancy rate over the dataset average.

A study by Ziebe et al. (1997) examines the effects of morphological components as defined by Cummins in 1986 (quality score and cleavage stage) on implantation rate. They focus on 3 clinics over a 3 year timeframe resulting in 1001 transfers of 1918 embryos. These clinics transferred embryos on day 2 and found 2 notable conclusions:

(1) Implantation rates were significantly higher for 4 cell embryos (23%) than 3 cell (7%) or 2 cell (12%); and (2) Significance in the difference between the IR of the best and worst morphology groups. It would be interesting to see the how the differences between each morphology grouping are presented when a larger sample size is examined.

Another attempt at predicting the results of IVF procedures focused on both zygote scoring and embryo morphology as determining factors for IVF success (De Placido, 2002). This research found minimal correlation between zygote scoring and resulting embryo morphology (r =0.1). With 15% of the poorest quality zygotes producing viable embryos, it is not strong enough to suggest that a low quality zygote can be thrown out. Additionally, minimal impact on pregnancy or implantation rates was shown in comparing good and poor quality zygotes. The key takeaway here is that embryo morphology is much more predictive than zygote score. Their results, albeit from a small sample size (183 patients) confirm Cummins's 1986 analysis of a correlation between EDR and embryo success, and suggest that this is more powerful than looking at zygote or embryo morphology scores alone. They also mentioned that the presence of one good quality embryo was more important than the mean embryo score.

Another proposal was generated to score embryos on a 4 point scoring method for day 2 embryos. This study focused on 957 single embryo transfers (Giorgetti et al., 1995). Looking only at single embryo transfer cases is an interesting limitation. The benefit is that it eliminates any confounding effects that could collectively affect embryos in a multiple embryo transfer. Perhaps more importantly, it removes the inherent uncertainty that results from cases such as: two non-identical embryos are transferred and one is successful. It would be nearly impossible and certainly very costly to track which of the embryos implants. The drawback is that this subset of patients consists of not only elective single embryo transfers but also patients in which only one viable embryo was available. This latter effect could cause some confounding effect of less viable patients compared to the overall pool. The researchers found several significant factors that affected pregnancy rate. One was rate of development: 4 cell embryos implanted twice as well as slower or faster growing embryos. Also, they found that in 99 cases of uncleaved (1 cell) embryos, 3 pregnancies and 0 live births. Based on this a 4 point scale was derived with one point given for each of the following:

1. Cleaved embryo presence
2. Minimal or no fragmentation
3. No irregular blastomeres
4. 4-cell stage

A significant correlation was shown between the embryo scores and both pregnancy rate and take home baby rate. Interestingly, the researchers did not show any difference in embryo score between younger patients and older patients (age > 38).

Each of the components of this score is individually significant in terms of their effect on pregnancy rate. However, each effect is not equivalent in magnitude. The authors are not clear as to the criteria they used to decide on the weighting.

A major prerequisite in achieving the goal of reducing multiple pregnancies and shifting towards single embryo transfer is being able to identify embryos with a high implantation potential. The best way to know more about the potential of any embryo is to grow it for longer i.e. wait until day 5 to transfer instead of day 3. The issue with this is that culturing practices are not perfect; there are fewer embryos suitable to transfer, leading many clinics to transfer earlier. (Royen et al., 1999) This raises the question, were the embryos that became unviable between day 3 and day 5 never going to have the potential to result in a healthy baby, or can this sometimes be a result of an exogenous factor from the laboratory?

Royen et al. (1999) believed the latter is true. This would indicate that a clinic can improve by eliminating these exogenous factors and improve their success by doing more day 5 transfers. Certainly it is not that simple, as it is very difficult to even identify what these factors may be, not to mention remediate them. This will require many future designed experiments and likely significant costs if the factors are precise control of temperature or air quality, for example.

Royen et al. (1999) took another approach; he said that if he can better identify which embryos are most viable earlier on in the process, the same embryo selection can be achieved with an earlier transfer. Thus, lower exposure to problems that could occur during the culturing.

To explore this, he conducted a study to try to identify the best criteria to evaluate embryo potential. He scored embryos based on (i) fragmentation (ii) number of blastomeres and (iii) number of multinucleated blastomeres and recorded the three criteria on day 2 and day 3 of each cycle. This sounds very promising to have such detailed information tracked. Unfortunately, Royen's sample size is quite small and his results have limited significance.

His study examines 400 cases over 2 years at a clinic in Belgium using relatively homogeneous drug treatment and transfer procedure. In order to characterize high quality embryos, they look at the 23 cases where 2 embryos are transferred and a dizygotic twin pregnancy occurs. In these 46 embryos, none of them were fragmented and the majority had 8 blastomeres on day 3.

Royen et al. (1999) claimed that they "preferred not to set an upper limit because the faster an embryo cleaves, the more likely it is to implant successfully." This is not consistent with other studies performed on larger sample sizes that indicate embryos that grow too fast have reduced performance (Cummins et al., 1986). It is also mentioned that no article to date has described the embryos with maximal implantation rate. Further, the relative potential of embryos of differing morphology combinations has not been explored extensively.

Light has been shed on the lack of a universal, reliable Day 3 scoring system since the early days of IVF. Desai (2000) considers several methods of scoring embryos in attempt to better predict their potential in resulting in a healthy baby. The 3 proposals are listed below:

1. Number of cells
2. Number of cells - 2 points if heavily fragmented
3. Number of cells - 2 points if heavily fragmented + 0.4 points for each positive feature: (Blastomere expansion, All equal size, Absence vacuoles, Pitting, Compaction)

Desai and her associates considered 316 embryos from 93 patients to test the validity of the scoring methods. In each of the 3 methods, the score obtained from the pregnant group was significantly higher than the not pregnant group. But is this significant difference enough to provide useful predictive information or relative value of these embryos? Desai also looks at the actual average implementation rate of each category.

The authors of this study point out that in many cases, they have several day 3 embryos with similar cell counts and levels of fragmentation and need additional ways to determine the best embryo to transfer. Hence, they derived the 5 morphological parameters added to their score. Based on their data, they are that these morphological parameters have greatly improved their ability to decide on the optimal embryos to transfer. The challenge will be to standardize the tracking of these parameters in a large dataset across multiple clinics to see if the results are confirmed.

Interestingly, the 5 morphological predictors were not significant on their own in predicting outcome, however when combined in the score they were. This could be due to the fact that the number of cells was by far the largest component of the score, and number of cells is very significant.

The drastic improvement in the 1990's of drug treatments in ART technologies and the resultant ability to produce more eggs and embryos has been well documented (Hu et al., 1998). As embryos are more plentiful, it becomes increasingly important to understand how to score embryos absolute and relative potential. More effective scoring will lead to both higher instances of singleton pregnancy and lower rates of unwanted multiple pregnancies.

> A scoring system is proposed as a 1-5 grading creating groupings based on combinations of number of sells, equivalence of blastomere size and fragmentation. This grade is then subtracted from 5 to obtain an embryo scoring. Table 1 shows the breakdown:

Table 1: Embryo Grading and Score System

Grade	Score	Morphology
1	4	≥5 cells; blastomeres of equal size; 0 cytoplasmic fragments
2	3	≥5 cells; blastomeres of equal size; ,30% cytoplasmic fragments
3	2	≥5 cells; blastomeres of distinctly unequal size; 0 cytoplasmic fragments
4	1	≥5 cells; blastomeres of equal or unequal size; 30%–50% cytoplasmic fragments if equal or 1%–50% fragments if unequal
5	0	<5 cells of any size, or any pre-embryo with .50% cytoplasmic fragments

Adapted from Hu et al. (1998) with permission

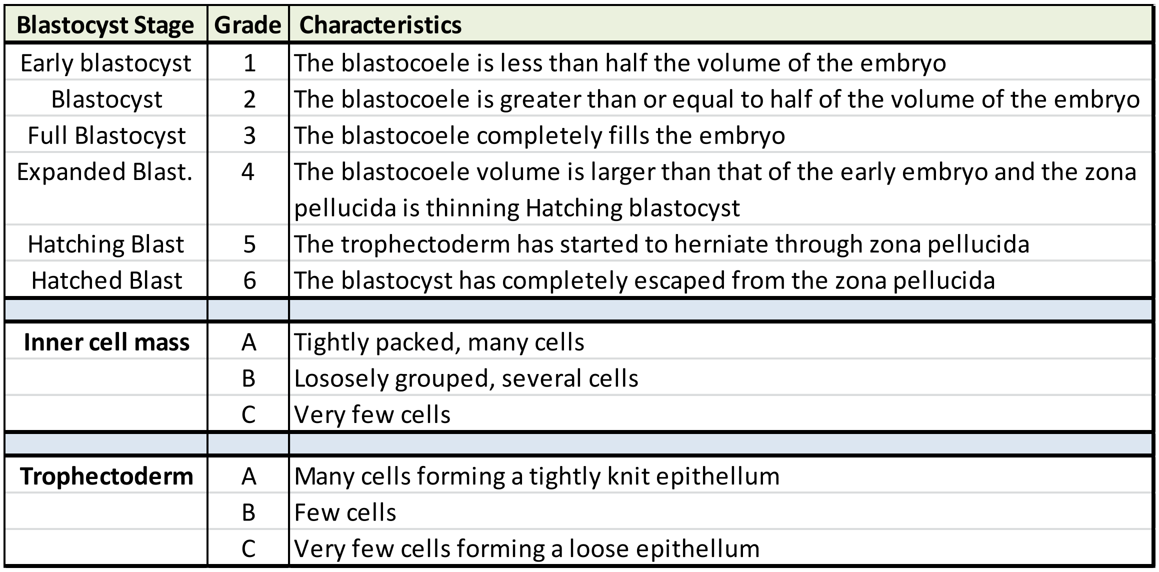
Then, Hu added the mean embryo scores for each case to create a "mean cumulative embryo score" for all of the embryos transferred in each cycle. Perhaps "cumulative mean embryo score" would be a more intuitive naming scheme. The issue with using a cumulative scoring for all embryos transferred is it provides a much higher score for cases where more embryos were transferred. Further, adding scores together makes the assumption that the probability of implantation of one embryo is independent of the probability of another (possibly identical) embryo implanting in the same uterus.

Regardless, 754 consecutive patient cycles were scored. The patients were split up into 3 age groupings (<36, 36-39, >39). Also the MCES was split up into (<10, 10-19, >19). In every case, there was an upward trend of implantation, pregnancy and multiple pregnancy rate. The MCES is an attempt at assessing quality and number of embryos transferred in one metric. Hu (1998) shows that higher-order multiple conceptions are more prevalent in higher quality embryos in younger patients as well as obviously in cases where more embryos are transferred. He goes on to suggest a recommended number of embryos to transfer in each case. In patients under 36, he suggests transferring up to 4 poor quality, 2 fair quality, or 2 good quality embryos.

A key discussion centers around the evaluation criteria for determining the best embryo(s) to transfer. Suggestion has been made that in many cases the selection is based more on clinical tradition than scientific evidence (Matchinger 2013). Matchinger mentions a key challenge of making accurate assessments of embryo evaluation methodologies due to the challenge of assessing a large enough homogeneous sample. Either the sample will be limited in size when coming from a particular clinic, or it will be biased by differing patient selection or evaluation criteria by different clinics. Every embryo is unique and the line that one doctor draws between a moderately and severely asymmetric embryo may not be identical to another.

Blastocysts (typically found in day 5 transfers) are generally scored by the method proposed by Gardner et al. (2000). Gardner's system considers the stage, the inner cell mass and trophectoderm. His score and has been shown to be more effective in predicting IR even than more recently developed methods. From this, SART has developed a standardized embryo scoring method; the complete array of levels for each descriptor is shown in Table 2:

Table 2: Gardner's System for Scoring Blastocysts



Blastocyst Stage	Grade	Characteristics
Early blastocyst	1	The blastocoele is less than half the volume of the embryo
Blastocyst	2	The blastocoele is greater than or equal to half of the volume of the embryo
Full Blastocyst	3	The blastocoele completely fills the embryo
Expanded Blast.	4	The blastocoele volume is larger than that of the early embryo and the zona pellucida is thinning Hatching blastocyst
Hatching Blast	5	The trophectoderm has started to herniate through zona pellucida
Hatched Blast	6	The blastocyst has completely escaped from the zona pellucida
Inner cell mass	A	Tightly packed, many cells
	B	Lososely grouped, several cells
	C	Very few cells
Trophectoderm	A	Many cells forming a tightly knit epithellum
	B	Few cells
	C	Very few cells forming a loose epithellum

Adapted from Matchinger et al. (2013) with permission

Even with a standardized scoring system, there is a degree of intra-observer and inter-observer variability in scoring. Some studies have focused on measuring this such as (Paternot et al., 2011a) which had 5 embryologists look at the same embryos and compared their assessment. Variation was deemed relatively low, but the fact that some exists is demonstrative of the challenge in developing standardized criteria for evaluating day 3 embryos.

The primary effort of formulated embryo scoring considering multiple criteria together was by Cummins' 1986 study, however these results are reflective of a very early stage in the process. Matchinger mentions that more work needs to be done to revise these results and consider them in the context of a larger dataset and modern practice techniques.

Dr. Gibbons (2007) discussed why it is necessary to make improvements in the predictability of IVF cycles. One area where this is vital is surrounding the key decision of the IVF process: which and how many embryos to transfer. Thus, it is necessary to understand the factors that can affect the performance of an embryo. As described in detail in an earlier section, the morphological criteria examined for each type of transfer are primarily number of cells, symmetry and fragmentation for day 3 embryos. For day 5 embryos, stage (hatching, expanded, or early blastocyst), inner cell something and trophectoderm morphology.

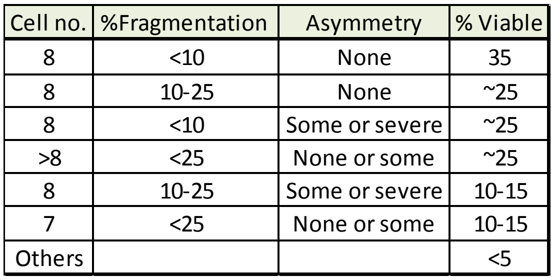
The largest study on exploring the impact of these morphological predictors is Dr. Racowsky's 2003 study, which looks at a SART dataset of 5,112 patients from 1998 to 2001. This study brings up a very important point: the relative value of day 3 morphological combinations and embryo viability remain ill defined. The last attempt at establishing an embryo score was by J.M. Cummins (1986).

Quite logically, Racowsky and her colleagues focus on the output of implantation rate measured by week 8 fetal heart rates rather than pregnancy rates. They argue that problems that occur past week 8 are more likely due to biological complications not related to embryo quality. Also, in addressing the same issue that Cummins faced, of not knowing which embryo implants in the case of multiple embryo transfer, they used a fairly innovative strategy: consider all of the cases in which the embryo's fate is known:

1. Either all or none of the embryos implanted
2. All of the embryos transferred are morphologically equivalent

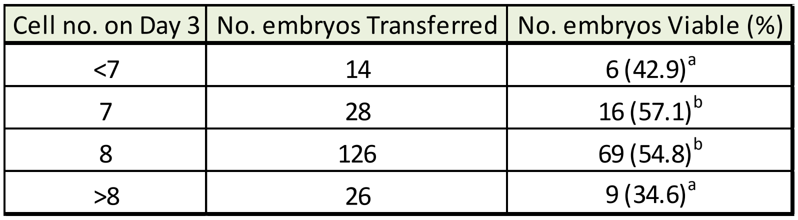
The researchers looked at a group of 1,823 embryos transferred on day 3 all from patients younger than 37 years old. They considered the effects of fragmentation asymmetry and number of cells univariately and found significance in each case. They then proposed viability percentages shown in Table 3 – Table 5 below:

Table 3: Interaction of Conventional Day 3 Morphology Markers as Predictors of Viability



Cell no.	%Fragmentation	Asymmetry	% Viable
8	<10	None	35
8	10-25	None	~25
8	<10	Some or severe	~25
>8	<25	None or some	~25
8	10-25	Some or severe	10-15
7	<25	None or some	10-15
Others			<5

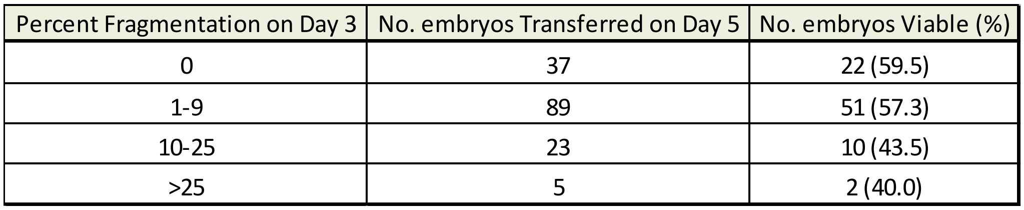
Table 4: Cleavage Stage on Day 3 and Viability of Expanding/Expanded Blastocysts on Day 5



Cell no. on Day 3	No. embryos Transferred	No. embryos Viable (%)
<7	14	6 (42.9)[a]
7	28	16 (57.1)[b]
8	126	69 (54.8)[b]
>8	26	9 (34.6)[a]

Different superscripts indicate a significant difference, a versus b, $p < .04$.

Table 5: Fragmentation on Day 3 and Viability of Blastocysts on Day 5



Percent Fragmentation on Day 3	No. embryos Transferred on Day 5	No. embryos Viable (%)
0	37	22 (59.5)
1-9	89	51 (57.3)
10-25	23	10 (43.5)
>25	5	2 (40.0)

No significance difference was found between the groups.

Tables 3 – 5 adapted from Racowsky et al. (2003) with permission.

They also studied some day 5 embryos but the sample size was quite small (258 embryos). This study was performed based on data from one clinic and provides an important foundation for the methodology of assessing embryo quality based on morphological predictors. This study is fairly similar to ours but it uses a much smaller dataset.

It is evident from the research that has been done that there is interest in better understanding the impacts that embryo morphology has on treatment outcomes. Several researchers have sought to develop a universally applicable embryo score. The most robust methodology seems to be that of Racowsky et al. (2003) which derives a score directly from performance.

Advancements in data tracking facilitated by SART, among others, have allowed for larger sets of data to be available for analysis. This thesis will employ a methodology similar to Racowsky et al. to data from 40 clinics in an attempt to understand the value of embryos of various morphologies. These values will be comparable in one-dimension and allow clinics to track their performance accurately.

CHAPTER 3: METHODOLOGY

3.1 Exploration of SART-CORS Dataset

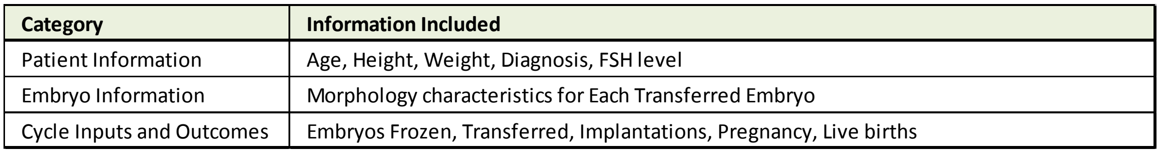
This retrospective study is based on a large dataset provided by the Society for Assisted Reproductive Technology (SART). The dataset obtained from SART contains information about 36836 cycles from 2009-2011. What distinguishes this dataset from many others that have been considered in previous studies is its size and its breadth.

Rather than coming from one or a couple of clinics, this data comes from 40 clinics across the United States. Since there is known to be considerable variability between the performance among clinics, it is important to ensure that the dataset sufficiently encompasses the population characteristics. To do this, the clinics were ranked by their live birth rate in patients under 35. Then 4 clinics were randomly selected from each decile of the ranking (i.e. 4 clinics between the 1st and 9th%ile, 4 from between the 10th and 19th %ile, and so on). The patient cases from these clinics represent the dataset upon which the conclusions of this study are based.

Due to doctor patient anonymity requirements, the individual clinics cannot be identified. Each cycle is given a unique identifier, but is not attributed to a particular clinic. The uniqueness of the dataset provides both opportunities and challenges. It is differentiated by the sheer quantity of data, which adds considerable power to the conclusions drawn. Also, it allows comparisons to be made between the best performing clinics and the poorer performing clinics.

The information included can be broken down into 3 main categories, shown in Table 6:

Table 6: Dataset Contents



Category	Information Included
Patient Information	Age, Height, Weight, Diagnosis, FSH level
Embryo Information	Morphology characteristics for Each Transferred Embryo
Cycle Inputs and Outcomes	Embryos Frozen, Transferred, Implantations, Pregnancy, Live births

All of this data can be overwhelming at first, with dozens of columns for each of the 36,836 entries. It was necessary to cater the data to the specific research objective of the study: comparison of clinics. The data is translated into usable information, which will improve understanding and ultimately drive actions by IVF practitioners in order to move in the direction of their goals: improving success rates and predictability of embryo transfers.

As is typically the case with such a large dataset, not every entry will be usable. Particularly, this is true when dealing with compiled data that is entered by many different people; each clinic's data reporting is not going to be perfect. Throughout the analysis, reductions will need to be made to maintain accuracy. It will be documented in each case when such reductions are made. The major adjustment that is initially made is removing the 6685 listed cases in which all embryos are cryopreserved (frozen) and 0 fresh embryos are transferred and. Since this study focuses on fresh embryo transfers, these cases are not relevant.

The first stage is using drug treatment to cause ovarian hyper stimulation and cause the woman to produce a high number of eggs during a particular menstrual cycle. Although there is significant amount of variability from one patient to the next, there doesn't seem to be a huge impact of clinic related factors on this stage. The overall average number of eggs produced in young patients (Age <36) is 14. As detailed in the chart on the next page, the avg. number of eggs retrieved from young patients (<36) does not significantly differ among clinics. Although there is a very slight upward trend, in each case, the range stays between 12 and 16 eggs. This is illustrated by Figure 2.

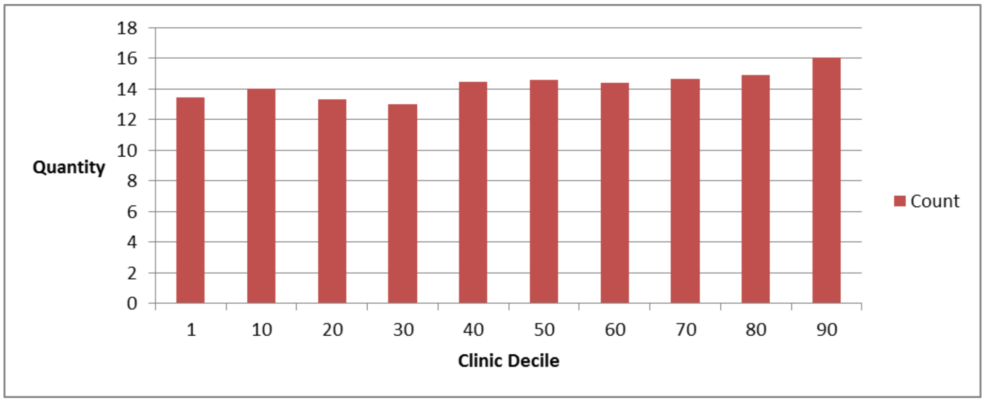


Figure 2: Average Oocytes Retreived in Patients <36

The stage of the process is where things really get interesting from a clinic perspective is growth of an egg into and embryos over the course of the first 3-5 days after fertilization. The reason this is important is the clinic is able to have a significant influence on the performance of this stage. These three to five days are the key differentiator between IVF from a traditional pregnancy. This short amount of time is extremely critical and can have a major impact on the success or failure of the treatment. Rather than inside of a woman's body which has evolved over millions of years to naturally regulate temperature and other factors that may influence the embryo's development, the young zygote is growing in a petri dish in a laboratory where it can be potentially exposed to many external factors such as impediments in the air or changes in temperature or humidity. These are merely speculative examples of factors that could potentially influence embryo success.

Since it is the portion of the process most controlled by the clinics, the early stage of embryo development pre-transfer is the most analogous part of the process to a traditional manufacturing process and the focus of this study. The clinics aim to control the output of this stage: viable embryos. The relevant measure of success from this stage is the Per Embryo Implantation Rate (PEIR). In the simplest terms possible, PEIR measures the portion of embryos transferred that result in a successful implantation.

$$PEIR = \frac{\#\ of\ implanted\ embryos}{\#\ of\ embryos\ transferred}$$

Since the objective of IVF clinics is to grow viable embryos, the most objective way to measure the viability of these embryos is by examining their real performance. Although it is quite commonly said in colloquially dialogue “you can’t be half pregnant”, growing humans is not as black and white as manufacturing metal castings, there are many treatment outcomes that can occur. As a result, the calculation must take into account many nuanced cases that are present and well documented in the SART dataset.

The PEIR calculation was made in the SART dataset using several columns of information available. The denominator is very straightforward; it is simply the quantity of fresh embryos transferred to the uterus on either day 3 or day 5.The numerator is much more complex and is calculated based on the following criteria shown in Table 7:

Table 7: Implantation Calculation

IF	Implantations =	Rationale
TreatmentOutcome_Not Pregnant (col AS) = Y	0	Pregnancy test was negative hence, no embryos implanted
TreatmentOutcome_Biochemical (col AT) = Y	1	Pregnancy test was positive, therefore AT LEAST one embryo implanted; however since nothing is visible on the ultrasound at 6 to 8 weeks, cannot know if more than one implanted, so by convention, a value of "1" is assigned
TreatmentOutcome_Ectopic (col AV) = Y	1	This is a tubal pregnancy; while it is possible to have more than one implant in the tubes it is very uncommon, so by convention a value of "1" is assigned

TreatmentOutcome_Heterotopic (col AW) = Y	UltrasoundFetalHearts +1	Very uncommon situation wherein there is both an implantation in the fallopian tube + implantation in the uterus; it would be extremely uncommon to have more than one in the tube, so we assign a value as above.
TreatmentOutcome_Clinical Intrauterine Gestation = Y And UltrasoundFetalHearts = 0	1	Early miscarriage; there is at least one implanted embryo
TreatmentOutcome_Clinical Intrauterine Gestation = Y And UltrasoundFetalHearts > 0	UltrasoundFetalHearts	Pregnancy

The formula used to calculate implantations in excel is shown in Appendix C.1.

On the other side of the spectrum is the final stage of the process in which the implanted embryo grows into a baby. This portion of the development is again very dependent on patient: it is happening quite literally inside of the patient. To verify this assumption, consider the translation between implantation rate, calculated as described above, and live birth per embryo transferred. As seen in Figure 3, the ratio of implantation rate to live birth remains relatively constant among clinics.

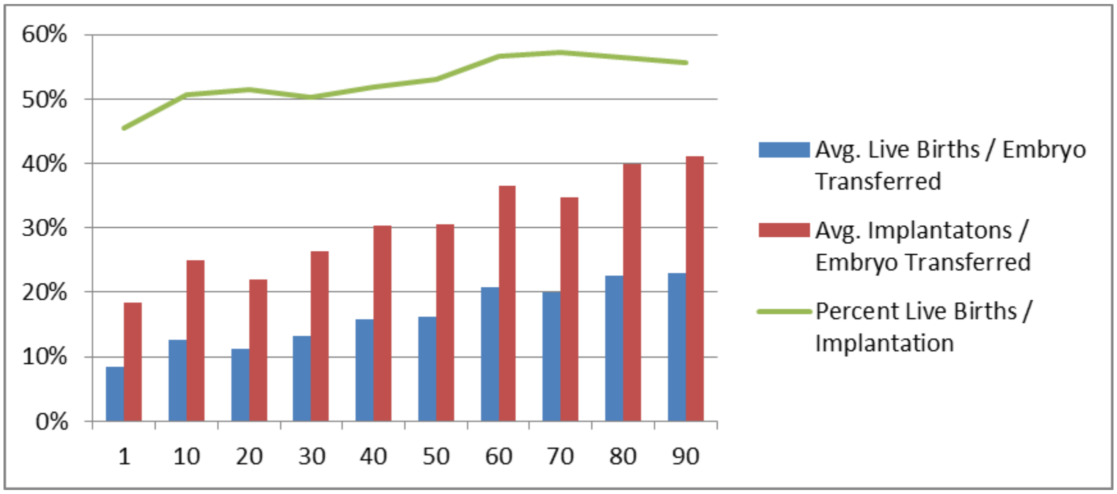


Figure 3: Implantation Rate to Live Births per Embryo Transferred Comparison

There has been significant debate as to what the most relevant output metric is to measure IVF clinic success (Min et al., 2004; Davies et al., 2004) and there debate is certainly merit to the arguments for considering singleton live birth rate as such. Singleton live birth is the goal of the IVF process so the portion of cycles that result in this is definitely an important metric when talking about clinic performance. However, PEIR is the direct outcome of the particular stage of the process within the control of clinics. Therefore, PEIR is the most relevant outcome and will be the metric of interest throughout this study.

Patient Effects

As the previous section indicates, there is inevitably a large amount of variability attributed to patients. Certainly, the input is much more variable when compared to typical manufacturing processes. Take, for example, a company that manufactures small precision cast iron parts. Every portion of the process is understood very well scientifically: the melting temperature of the iron, the strength of the mold cavity required, the hardening time, etc. Every piece is understood to a molecular level, and thus extremely high repeatability exists in the process and many high volume foundries report very low defect rates. To the contrary, IVF is a complex biological process with the outputs being live human beings. While vast knowledge is available in the biological and medical fields, it is well established that the human body is not understood as well as, say, molten metal. Even many parameters that are understood are impossible or at least very expensive and impractical to measure.

Given that the inputs will not be able to be perfectly measured or controlled, it is important to take advantage of any information that is readily available in order to understand the trends that exist. With this in mind, some patient factors have been found to be significant in predicting live birth rates in previous analyses such as age and maxFSH level (Ziebe et al., 2001; Frazier et al. 2004). Another study found the number of cryopreserved embryos to be significant, as this is an indication that the patient had a high number of viable embryos (Stern et al. 2009).

Since we have demonstrated that Implantation rate correlates to live birth rate in the previous section, it will be appropriate to consider implantation rate as the output of interest for the remainder of this analysis.

One of the first interesting things to look at in this exploratory study was the effects of the most obvious patient factor, age, and how prevalent it is in this dataset. Previous studies have overwhelmingly pointed to a decline in production of oocytes and high quality embryos the mid 30’s. In the SART data, these findings were echoed.

Figure 4 shows the trend of oocyte production by age from the women of various ages. The dataset contained patients from ages 19-52, however sample sizes were small (<25) outside of the 24-48 age range. An unexplained spike downward occurred for the group of 169 23-year-old patients. Overall, the trend was clearly downward.

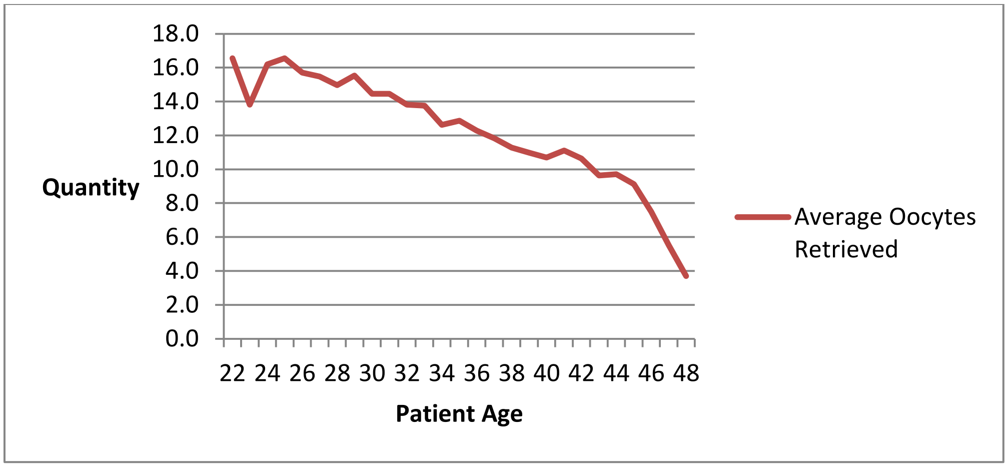


Figure 4: Average Oocytes Retrieved by Age

To measure the embryo quality from each age group, implantation rate was used. The average PEIR by age is shown here:

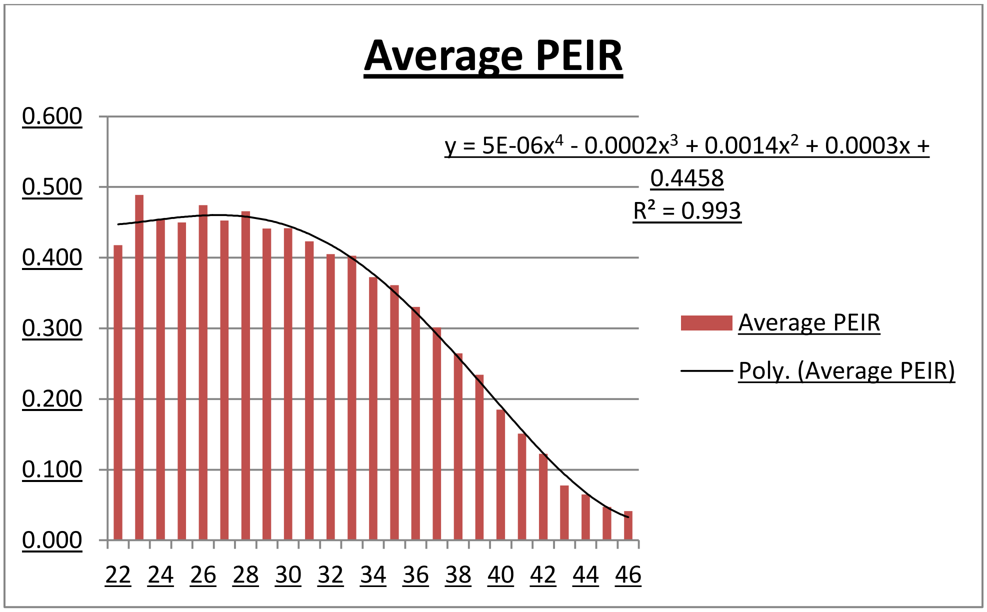


Figure 5: Average PEIR by Age

Initial exploration of the data reveals that the dataset is consistent with the findings of previous studies on effects of patient age. PEIR drops considerably as patients age. The average implantation rate can be modeled very accurately by the 4th degree polynomial shown in Figure 5.

Additionally, patient Body Mass Index (BMI) and smoker (Y/N) can be derived from the dataset. Each was considered and for significance independently. BMI is calculated from the given patient heights and weights. Patients are grouped by underweight, normal, overweight and obese. The mean PEIR is shown for each group in Figure 6.

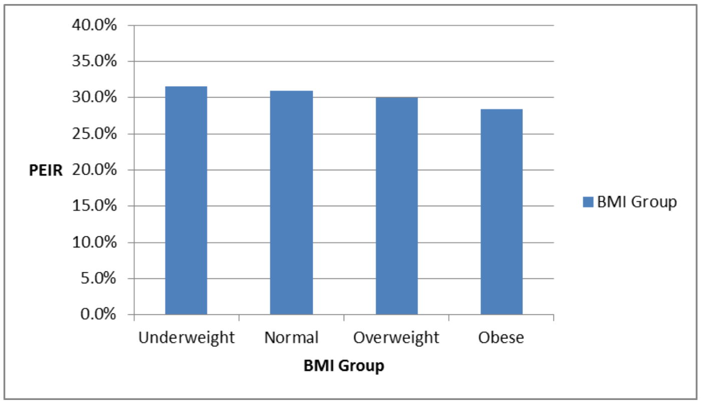


Figure 6: PEIR by Patient BMI

The comparison of PEIR by smoking status is shown in Figure 7. There is no practical difference between the groups, with smokers showing marginally higher PEIR on average.

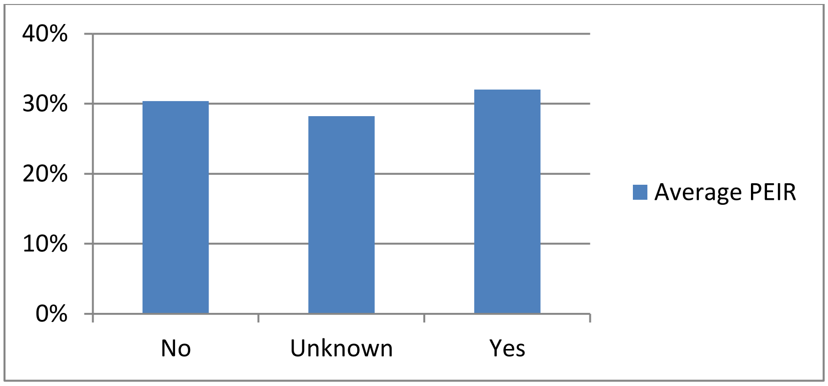


Figure 7: PEIR by Smoking Status

3.2 Development of Tools

One major goal of this paper is to propose a method for scoring day 3 and day 5 embryos derived directly from actual results. Developing this measurement of embryo quality is an essential prerequisite for making comparisons between clinics. Additionally, scoring embryos will enable practitioners to better assess their risk of twinning or HOM pregnancies given the number of embryos transferred.

The need to have a universal and reliable method of scoring embryos has been well established, and several attempts have been made in developing scoring systems

(Ziebe et al. 1997; Desai et al., 200) since Cummins' initial attempt in the 80s. The problem is that none of these systems are empirically derived from large datasets. In Cummins' case, and all other attempts up until Racowsky et al. (2013), it seems that the researchers followed a similar process: they examined the information that was available to them and found which factors were significant. Then, they converted values in each of these categories to a scoring system via a weighting. There are two main problems with this:

1. The relative weighting of factor effects is not effectively balanced.
2. It ignores potential interactions between variables (e.g. fragmentation and symmetry in day 3 embryos may have an interaction at various levels of each)

To avoid these problems, this analysis opts to omit the intermediate translational steps and instead derives the one-dimensional score directly from real results. As the previous section explains, the best measure of embryo performance is Per Embryo Implantation Rate (PEIR). This directly measures an embryo's capability to result in a pregnancy.

The descriptors reported in the SART dataset are different depending on the Embryo's Stage as detailed in .

Table 8: Embryo Descriptors in SART dataset

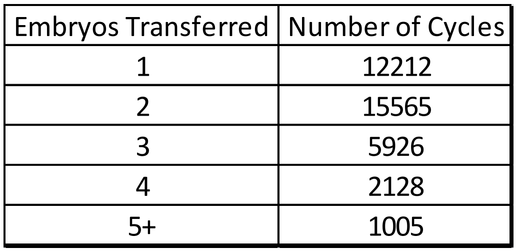
Stage	Reported Descriptors
Day 3	Number of Cells (1-8, >8), Symmetry (Perfect, Mod. Asymmetry, Severe Asymmetry), and Fragmentation (0, 1-10%, 11-25%, >25%)
Morula	Compaction (Complete, Incomplete), Fragmentation (0, 1-10%, 11-25%, >25%)
Blastocyst	Stage (Early, Expanded, Hatching), Inner-cell mass (Good, Fair, Poor), Trophoblast (Good, Fair, Poor)

Together these descriptors make up a morphology combination for each embryo. We generated a score for each particular morphology combination based on the average PEIR rate of that embryo type. For example, if 100 embryos are transferred on day 3 that are 8-cell embryo with moderate asymmetry and no fragmentation, and 26 implantations are resultant, this morphology combination would be assigned a score of 0.26.

A few obstacles exist in deriving this value from the dataset, so a several steps are required in making these calculations. The arranging and tallying of the data was performed in Microsoft Excel 2010. In the data provided by SART, each cycle represents a row, and up to 5 embryo's morphological characteristics are tracked. For each cycle, the number of embryos transferred is tracked and the number of implantations is calculated as described in the previous section.

The target is to draw conclusions about the performance of individual embryos, so the rows of data must be converted to embryos instead of cycle. Of the 36836 cycles contained in the dataset, the distribution of embryos transferred is shown in Table 9.

Table 9: Quantity of Embryos Transferred



Embryos Transferred	Number of Cycles
1	12212
2	15565
3	5926
4	2128
5+	1005

Thus, there exists morphological data for 74657 total embryos. From here a few reductions need to be made to reflect actual entries. First, there are 6685 cases where "NULL" values exist because this patient only had cryopreserved embryos and no fresh embryos were transferred. The next challenge that arises is that it is impossible to know which embryo successfully implanted in the case when multiple embryos were transferred. As previous researchers encountering this dilemma have done (Racowsky et al., 2003), the results will be derived from all cases when it can be known for sure the morphology of the embryos that implanted. These include the following situations:

1. No embryos implanted (e.g. 3 embryos are transferred and 0 implant)
2. All transferred embryos implant (e.g. 2 embryos are transferred and both implant)
3. All of the embryos transferred are morphologically identical (e.g. 2 Early Stage blastocysts are transferred with good inner cell mass and good trophoblast)

Although these seem like fairly specific cases, they actually span a good portion of the data available and fortunately maintain the integrity and power of the large dataset. The first two groups account for 36,693 embryos. After including all morphologically identical cycles (case 3), 46,267 remain with known implantation result.

The case of morphologically identical embryos is handled by giving each embryo "partial credit" equal to the portion of embryos that implant. In other words, if 4 identical embryos are transferred and 1 implants, each embryo is given credit for .25 implantations. The total of these four rows would then be 1 implantation for 4 embryos transferred.

From this very large sampling, the performance of a particular morphology combination can be shown. Developing a regression model was considered, however because many of the variables are categorical, the model would need nearly as many terms as there are groupings particularly if any interaction terms were included. It was most logical to instead develop contingency tables that calculate the PEIR for each treatment based on an average performance of all embryos of that type. As demonstrated in the earlier section, patient age has a very significant impact on implantation rate and separate

The data is brought into relational database software, Microsoft Access 2010, where it is grouped by each descriptor. An example of the query used can be found in Appendix C.2. The result is an average implantation rate for each embryo type. Table 10 shows the average implantation rate of each embryo combination for Day 3 Embryos in patients under 36. Table 11 and Table 12 show the other patient age groups. A confidence level is established for each group based on the sample size considered. The PEIR is a proportion therefore it follows the binomial distribution. The Standard Error can be estimated by the normal approximation as shown in Equation 2.

$$\begin{aligned}&p = proportion\ of\ transfers\ implanting\\&n = number\ of\ transfers\\&\alpha = level\ of\ significance = 0.05\\&s = number\ of\ successful\ implantations\end{aligned}$$

$$P = \hat{p} \pm Z_{1-\alpha/2} * SE \approx \hat{p}\ \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ Equation 1

Because some groups have small sample sizes, the Agresti-Coull method is used for slightly better accuracy. For the sake of consistency, Agresti-Coull intervals are used in all cases. Equations 2 and 3 show the adjustment.

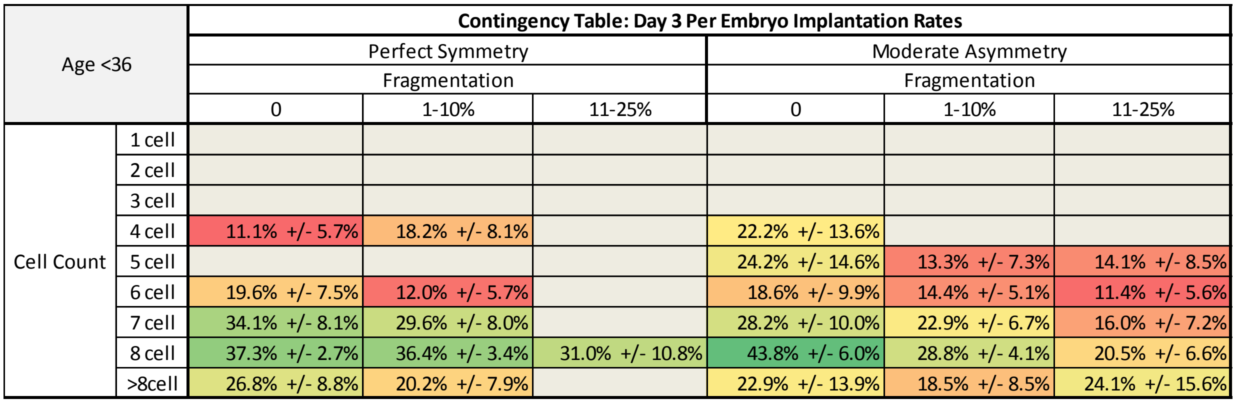
$$\tilde{p} \approx \frac{s+2}{n+4}\ \ ;\ \tilde{n} \approx n + 4$$ Equation 2

$$P = \tilde{p} \pm Z_{1-\alpha/2} * SE \approx \tilde{p}\ \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$$ Equation 3

A +/- 1.96 SE range is shown in the tables indicative of a 95% confidence interval. Groups with np<5 were not scored as the normal approximation is inaccurate. As a result, no severely asymmetric embryos were scored. From the entire dataset, 668 severely asymmetric embryos were transferred, resulting in 21 successful implantations, an IR of 3.1%. Perhaps, future studies can increase significance in these more uncommon groupings.

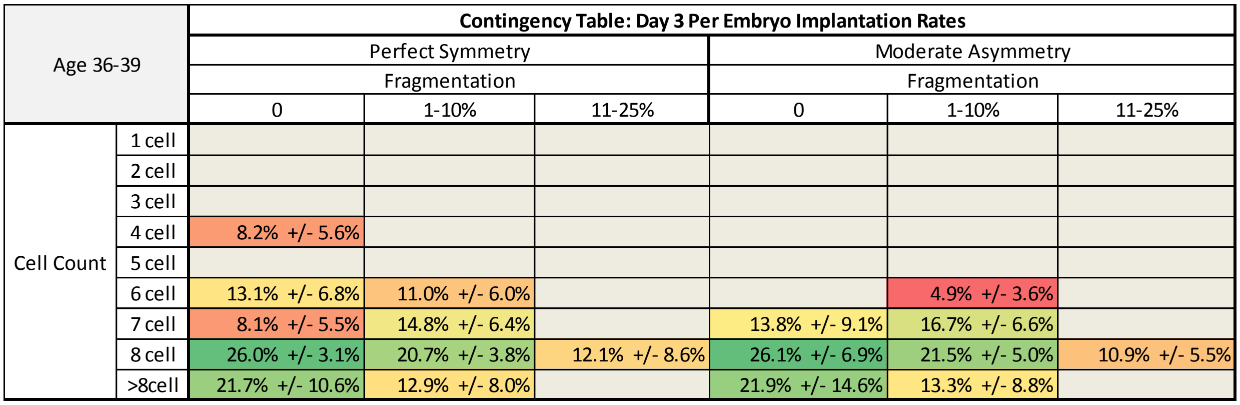
In the case of an IR of zero, the sample size is shown in parentheses in lieu of the confidence interval. See Appendix A for the sample size of embryos from all categories. Note that in some cases the Confidence bounds extend below 0. Implantation Rates are proportions and are of course bounded by 0 and 1.

Table 10: Day 3 PEIR and Confidence Intervals for Patients Age <36



Age <36		Contingency Table: Day 3 Per Embryo Implantation Rates					
		Perfect Symmetry			Moderate Asymmetry		
		Fragmentation			Fragmentation		
		0	1-10%	11-25%	0	1-10%	11-25%
Cell Count	1 cell						
	2 cell						
	3 cell						
	4 cell	11.1% +/- 5.7%	18.2% +/- 8.1%		22.2% +/- 13.6%		
	5 cell				24.2% +/- 14.6%	13.3% +/- 7.3%	14.1% +/- 8.5%
	6 cell	19.6% +/- 7.5%	12.0% +/- 5.7%		18.6% +/- 9.9%	14.4% +/- 5.1%	11.4% +/- 5.6%
	7 cell	34.1% +/- 8.1%	29.6% +/- 8.0%		28.2% +/- 10.0%	22.9% +/- 6.7%	16.0% +/- 7.2%
	8 cell	37.3% +/- 2.7%	36.4% +/- 3.4%	31.0% +/- 10.8%	43.8% +/- 6.0%	28.8% +/- 4.1%	20.5% +/- 6.6%
	>8cell	26.8% +/- 8.8%	20.2% +/- 7.9%		22.9% +/- 13.9%	18.5% +/- 8.5%	24.1% +/- 15.6%

Table 11: Day 3 PEIR and Confidence Intervals for Patients Age 36-39



Age 36-39		Contingency Table: Day 3 Per Embryo Implantation Rates					
		Perfect Symmetry			Moderate Asymmetry		
		Fragmentation			Fragmentation		
		0	1-10%	11-25%	0	1-10%	11-25%
Cell Count	1 cell						
	2 cell						
	3 cell						
	4 cell	8.2% +/- 5.6%					
	5 cell						
	6 cell	13.1% +/- 6.8%	11.0% +/- 6.0%			4.9% +/- 3.6%	
	7 cell	8.1% +/- 5.5%	14.8% +/- 6.4%		13.8% +/- 9.1%	16.7% +/- 6.6%	
	8 cell	26.0% +/- 3.1%	20.7% +/- 3.8%	12.1% +/- 8.6%	26.1% +/- 6.9%	21.5% +/- 5.0%	10.9% +/- 5.5%
	>8cell	21.7% +/- 10.6%	12.9% +/- 8.0%		21.9% +/- 14.6%	13.3% +/- 8.8%	

Table 12: Day 3 PEIR and Confidence Intervals for Patients Age >39

Age >39		Contingency Table: Day 3 Per Embryo Implantation Rates					
		Perfect Symmetry			Moderate Asymmetry		
		Fragmentation			Fragmentation		
		0	1-10%	11-25%	0	1-10%	11-25%
Cell Count	1 cell						
	2 cell						
	3 cell						
	4 cell						
	5 cell						
	6 cell				8.3% +/- 6.0%		
	7 cell				8.5% +/- 6.2%		
	8 cell	6.8% +/- 1.9%	4.8% +/- 1.9%			5.9% +/- 2.8%	5.1% +/- 3.7%
	>8cell						

The tables are color coded to make the trends more visible, with green reflecting the highest implantation rate and red reflecting the lowest implantation rates. Interestingly in patients under 35, the overall implantation rate is highest for moderately asymmetric 8-cell unfragmented embryos (n=256) over those with perfect symmetry (n=1188).

Day 5 embryos can come in two flavors: Morulas and Blastocysts. Each are scored using different criteria as described in Table 13 and Table 15 below detail the mean PEIRs and 95% confidence intervals for Morulas, using the same calculation methodology as above.

Table 13: Morula Contingency Table Age <36

Age <36	Compaction			Incomplete Compaction		
	Fragmentation			Fragmentation		
	0	1-10%	11-25%	0	1-10%	11-25%
Morula	13.9% +/- 2.8%	19.4% +/- 4.0%		18.6% +/- 5.0%	23.9% +/- 5.1%	14.4% +/- 5.3%

Table 14: Morula Contingency Table Age 36 – 39

Age 36-39	Compaction			Incomplete Compaction		
	Fragmentation			Fragmentation		
	0	1-10%	11-25%	0	1-10%	11-25%
Morula	25.4% +/- 4.6%	11.7% +/- 4.1%	26.9% +/- 8.7%	17.4% +/- 5.8%		

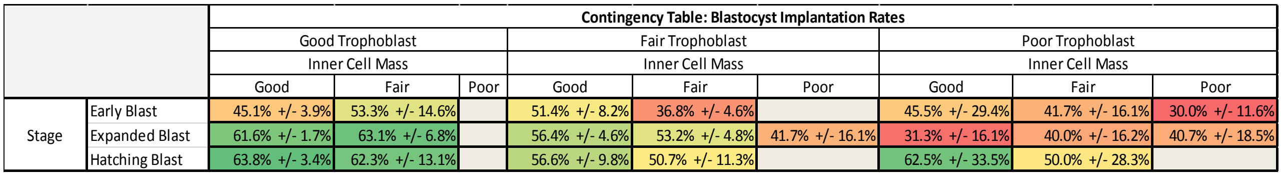
Table 15: Morula Contingency Table Age >39

Age >39	Compaction			Incomplete Compaction		
	Fragmentation			Fragmentation		
	0	1-10%	11-25%	0	1-10%	11-25%
Morula	17.6% +/- 5.2%					

The Morula tables were somewhat inconclusive due to low sample sizes. This is evident by the lack of clear trends and high standard deviations. Nonetheless, it is a start at representation of embryo performance.

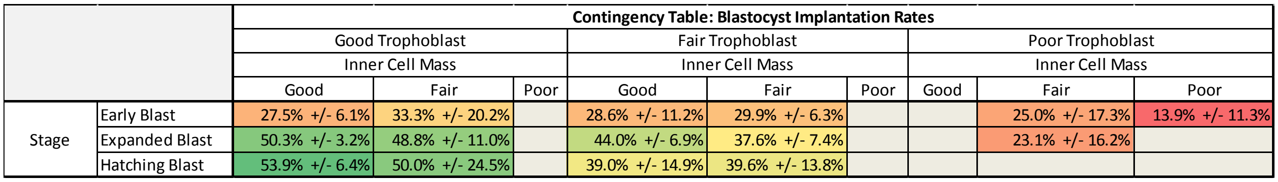
Blastocysts represent the most developed embryos that are transferred on day 5. On average, they are the most viable type of embryo transferred. The performance of each blastocyst morphology group is shown in the Table 16Table 18 below.

Table 16: Blastocyst Contingency Table Age <36



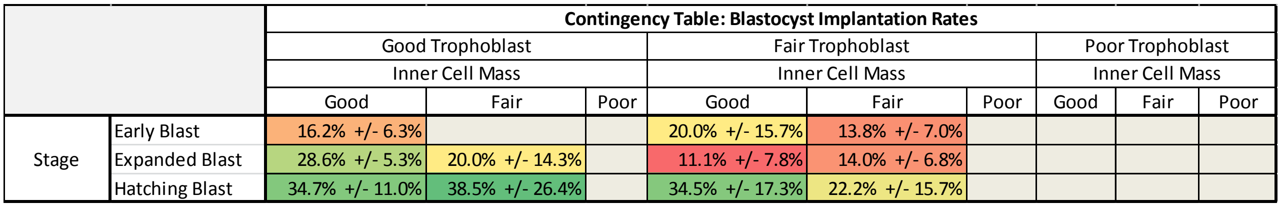
		Contingency Table: Blastocyst Implantation Rates								
		Good Trophoblast			Fair Trophoblast			Poor Trophoblast		
		Inner Cell Mass			Inner Cell Mass			Inner Cell Mass		
		Good	Fair	Poor	Good	Fair	Poor	Good	Fair	Poor
Stage	Early Blast	45.1% +/- 3.9%	53.3% +/- 14.6%		51.4% +/- 8.2%	36.8% +/- 4.6%		45.5% +/- 29.4%	41.7% +/- 16.1%	30.0% +/- 11.6%
	Expanded Blast	61.6% +/- 1.7%	63.1% +/- 6.8%		56.4% +/- 4.6%	53.2% +/- 4.8%	41.7% +/- 16.1%	31.3% +/- 16.1%	40.0% +/- 16.2%	40.7% +/- 18.5%
	Hatching Blast	63.8% +/- 3.4%	62.3% +/- 13.1%		56.6% +/- 9.8%	50.7% +/- 11.3%		62.5% +/- 33.5%	50.0% +/- 28.3%	

Table 17: Blastocyst Contingency Table Age 36-39



		Contingency Table: Blastocyst Implantation Rates								
		Good Trophoblast			Fair Trophoblast			Poor Trophoblast		
		Inner Cell Mass			Inner Cell Mass			Inner Cell Mass		
		Good	Fair	Poor	Good	Fair	Poor	Good	Fair	Poor
Stage	Early Blast	27.5% +/- 6.1%	33.3% +/- 20.2%		28.6% +/- 11.2%	29.9% +/- 6.3%			25.0% +/- 17.3%	13.9% +/- 11.3%
	Expanded Blast	50.3% +/- 3.2%	48.8% +/- 11.0%		44.0% +/- 6.9%	37.6% +/- 7.4%			23.1% +/- 16.2%	
	Hatching Blast	53.9% +/- 6.4%	50.0% +/- 24.5%		39.0% +/- 14.9%	39.6% +/- 13.8%				

Table 18: Blastocyst Contingency Table Age >39



		Contingency Table: Blastocyst Implantation Rates								
		Good Trophoblast			Fair Trophoblast			Poor Trophoblast		
		Inner Cell Mass			Inner Cell Mass			Inner Cell Mass		
		Good	Fair	Poor	Good	Fair	Poor	Good	Fair	Poor
Stage	Early Blast	16.2% +/- 6.3%			20.0% +/- 15.7%	13.8% +/- 7.0%				
	Expanded Blast	28.6% +/- 5.3%	20.0% +/- 14.3%		11.1% +/- 7.8%	14.0% +/- 6.8%				
	Hatching Blast	34.7% +/- 11.0%	38.5% +/- 26.4%		34.5% +/- 17.3%	22.2% +/- 15.7%				

In contrast to the Morula tables, the blastocyst tables show very clear and predictable trends within each age grouping. Almost universally, the average performance increases toward the lower left corner of the chart indicating that more developed blastocysts with better inner cell mass and trophoblast are better performing.

These contingency tables show with indicated confidence the performance that can be expected from each embryo morphology combination. This is a one-dimensional metric of quality.

CHAPTER 4: ANALYSIS AND RESULTS

Now that an empirically derived and accurate embryo scoring methodology has been developed, it can be used for a variety of applications. One such application is making comparisons in the ability of clinics to grow viable embryos. Growing morphologically superior embryos will ultimately result in better live birth rates and reduce the number of embryos needed to be transferred per patient. Morphologically superior embryos are defined as those that are historically more viable as illustrated by the contingency tables. Naturally, these are goals that all clinics strive for.

From the contingency tables, embryos are assigned a one dimensional score based on mean performance. For example, *6-cell, perfectly symmetric, nonfragmented* embryos transferred on day 3 have resulted in 19 implantations over 103 transfers, so the score for that embryo in patients under 36 is: 0.184. An embryo's viability can be effectively measured by this empirically derived score.

The confidence in each score is based on the number of embryos of that particular morphology considered in the dataset. Figure 8 provides a visual representation of the contingency table. A 95% confidence interval is shown for each group's average PEIR. The remaining patient age groups and the charts for blastocysts can be found in Appendix D.

In the context of comparing clinics, it is important to note that the "same embryo" e.g. *Perfect, 8-cell, nonfragmented, Day 3* transfer from a 40 year old patient is not as viable as an embryo from a 30 year old with the same descriptive values. Accordingly, age is also factored into the groupings in the contingency tables.

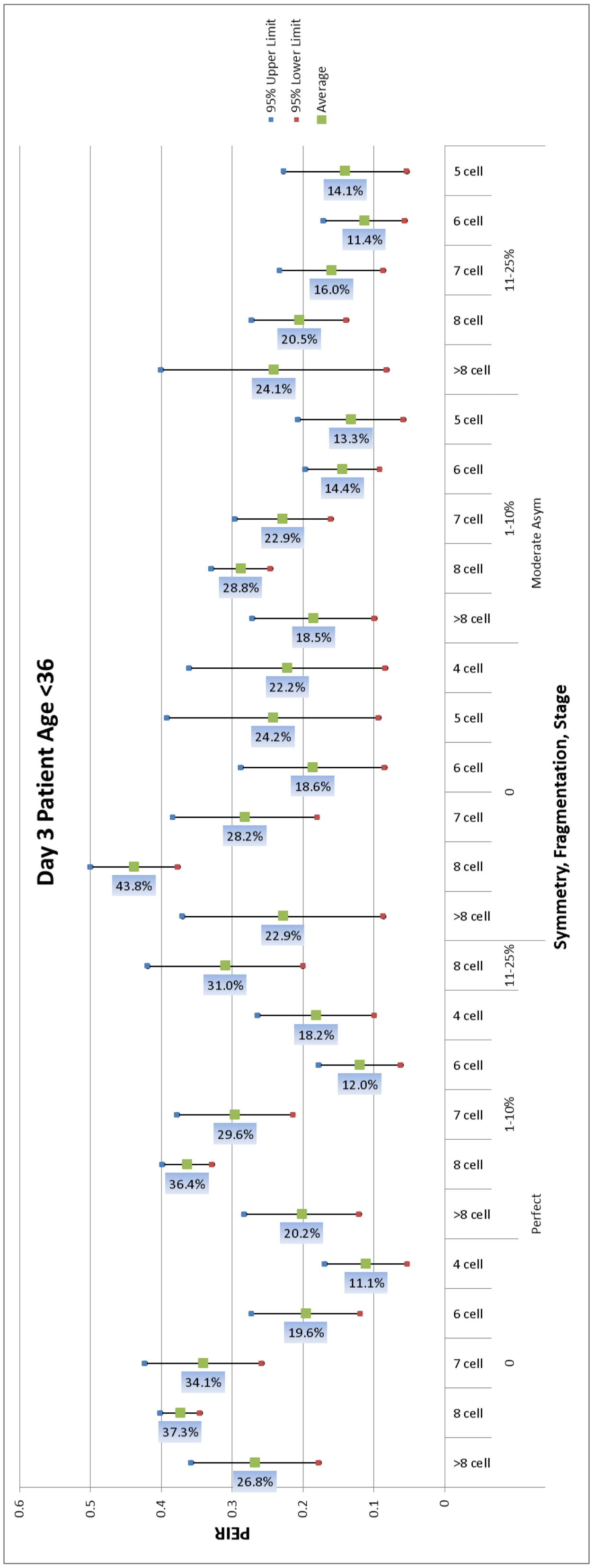


Figure 8: PEIR Confidence Intervals for Day 3 Emrbyos, Patient Age < 36

Once the scores have been established for each grouping, the first logical comparison of clinics is: how viable are the embryos that they are producing? The SART data collected from 40 clinics over 3 years was used to develop the contingency tables which assign scores to all embryos with sufficient sample size to be confident in the value. Starting with the 46,266 embryos transferred, each is matched with their embryo grouping, be it a Day 3 cleavage stage embryo, Morula, or Blastocyst. A good portion of these embryos did not contain complete information in the form of having "Unknown" or "Not entered" values in some columns. Also, those that were aligned with combinations that did not have sufficient sample size to produce a standard deviation of less than 10% were not included. 23,676 embryos remained for clinic comparisons.

The mean embryo score for each clinic decile is shown in Figure 9:

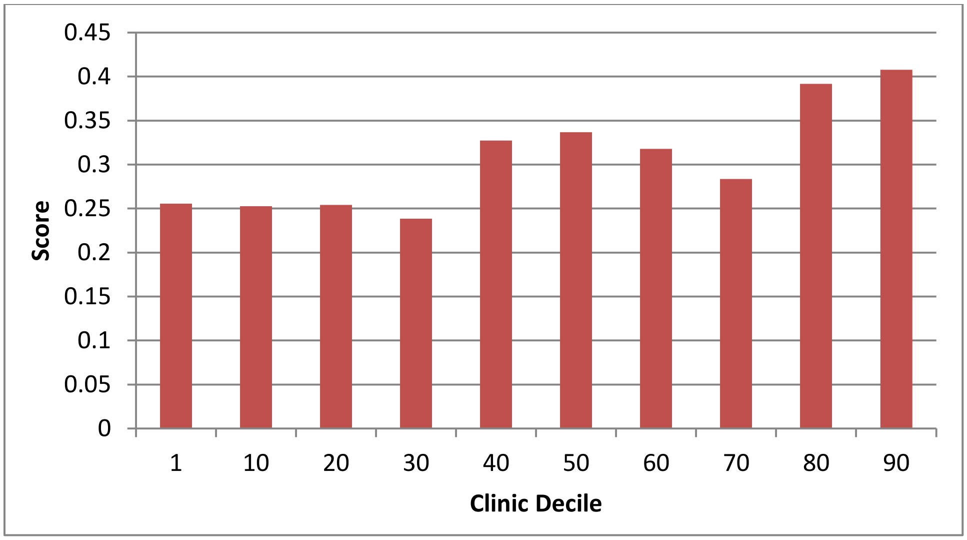


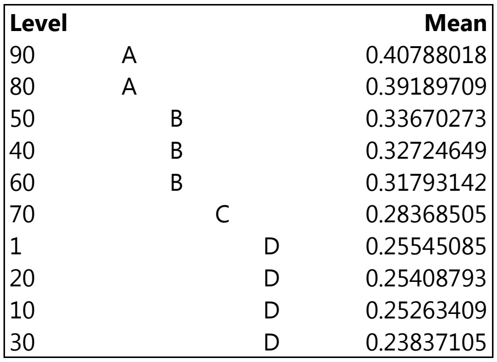
Figure 9: Mean Implantation Rate by Clinic Decile

An ANOVA was performed in JMP 11.0 and a highly significant difference was found between the clinic mean embryo scores (p<.001). Table 19 shows the ANOVA table.

Table 19: ANOVA of Mean Embryo Score by Clinic

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Percentile Group	9	65.3200	7.25778	162.2853	<.0001*
Error	23666	1058.3995	0.04472		
C. Total	23675	1123.7195			

Statistically significant differences exist between the groups not sharing the same letter in the Tukey-Kramer HSD comparison in Figure 10 below.



Level					Mean
90	A				0.40788018
80	A				0.39189709
50		B			0.33670273
40		B			0.32724649
60		B			0.31793142
70			C		0.28368505
1				D	0.25545085
20				D	0.25408793
10				D	0.25263409
30				D	0.23837105

Figure 10: Connecting Letters Report

The top two deciles show significantly higher embryo scores than the rest of the group. The 70th percentile group appears to be the only group that doesn't logically parallel the trend of live birth rate. This is because the 70th percentile group has the largest portion of patients over age 35. Overall though, the upward trend is clear.

It is evident that the better performing clinics (80th and 90th percentile) produce more viable embryos on average when compared to the lower performing clinics deciles.

One potential cause for this difference in score is that Day 5 transfers have over 2.5x the average performance compared to Day 3 transfers. The mean embryo score for Day 3 transfers is 16.4% compared to 43.6% for Day 5 transfers. The higher performing clinics make a larger portion of their transfers on Day 5 (Figure 11).

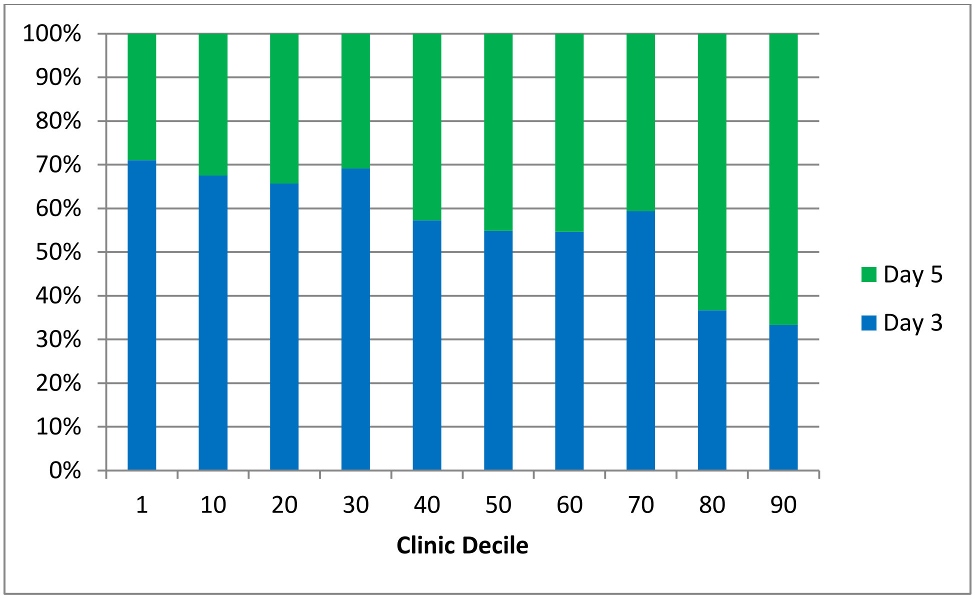


Figure 11: Day of Transfer by Clinic

Surely, a good portion of the difference is coming from the higher ratio of Day 5 to Day 3 transfers. An initial reaction to this might be to ask, why not simply transfer more embryos on day 5? One thing to keep in mind is that the reason why clinics transfer on Day 5 instead of Day 3. If the doctor plans to transfer two embryos in this patient and only two live embryos are remaining on Day 3, they will transfer those two embryos at that time. In another case, the conditions may be more favorable and there could be 5 good quality embryos available on day 3. In the latter case it makes sense to wait until day 5 and see which embryos develop the best, and at which point transfer the best embryos. Waiting until day 5 gives the doctor more information about how normally an embryo will develop. So, making a higher portion of transfers on day 5 is less of a decision and more of an indication of the clinic's ability to grow multiple good embryos.

Nonetheless, it is interesting to consider individually whether the better clinics are producing morphologically superior embryos within a day of transfer. Looking at Day 3 embryos only, the mean embryo scores by clinic are shown below (Figure 12). There does not appear to be any significant trend in the data.

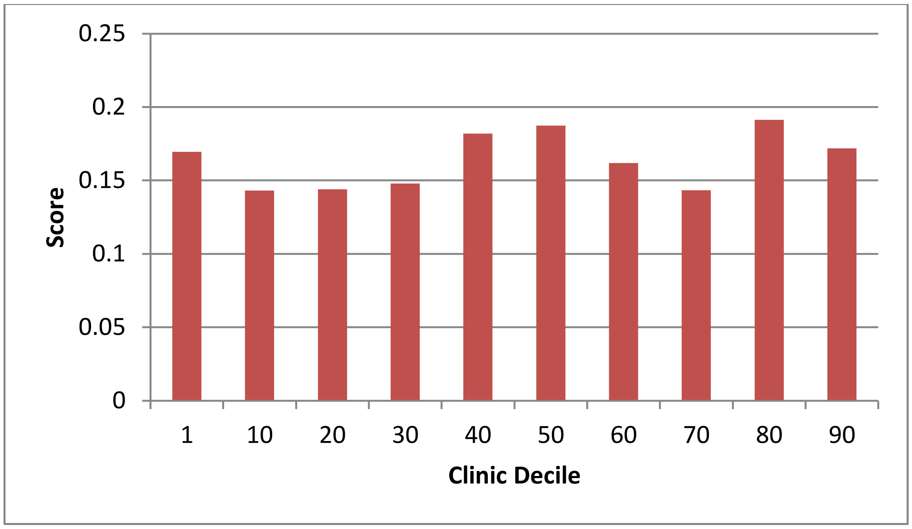


Figure 12: Day 3 Mean Embryo Score by Clinic Decile

The analysis is repeated for Day 5 (Figure 13).

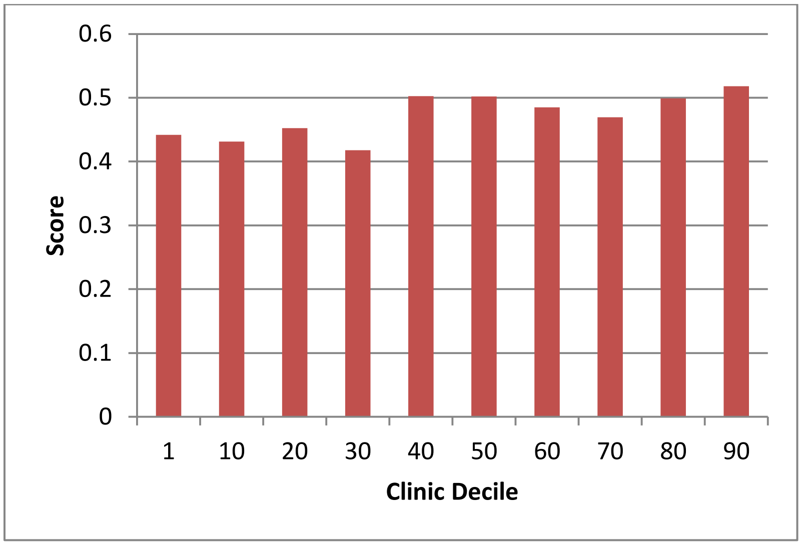


Figure 13: Day 5 Mean Embryo Score by Clinic Decile

While the difference is not extremely pronounced, a Tukey-Kramer comparison reveals that a significant difference exists between the lower 4 and the top 6 deciles with the exception of the 70th and 20th percentile groups showing no significant difference. The JMP output can be found in Appendix B. This indicates that the better performing clinics produce significantly more viable Day 5 embryos on average when compared to the lower performing clinics.

The high performing clinics are producing morphologically superior embryos on average. The next question that arises is: do morphologically equivalent embryos result in higher implantation rates at the best clinics? There are several ways to approach this; one would be to look within individual groups. The largest groups from each day of transfer are described in Table 20 below.

Table 20: Most Plentiful Embryo Combinations

Age	Morphology	Embryo Score	Quantity
<36	Expanded Blast, Good, Good	0.616	3273
<36	8 cell , 0 Frag. ,Perfect Sym.	0.373	1188

In each case, the performance of this particular embryo type is compared among the clinics. Results are as follows:

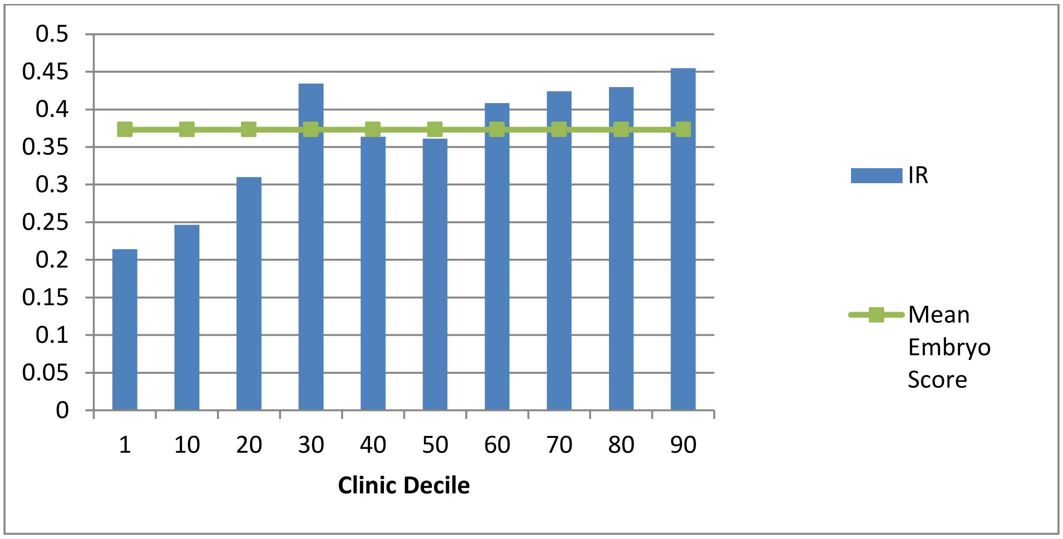


Figure 14: Performance of 8-cell, Symmetric, Nonfragmented Embryos by Clinic Decile

With the exception of the 30th decile, there is a very strong upward trend in the performance of the optimal Day 3 transferred embryos. The mean implantation rate of these embryos that have the exact same values based on SART's scoring categories is .455 in the top decile compared to .214 in the lowest, over a 2 fold increase. This has some very important implications when using the contingency tables for predictive value which are discussed in detail in the conclusion section.

Now, consider the embryo that is most abundant (n=3273): Blastocysts from young patients (<36) with good inner cell mass and trophoblastic scores transferred on Day 5. These show an overall average IR of 61.6%.

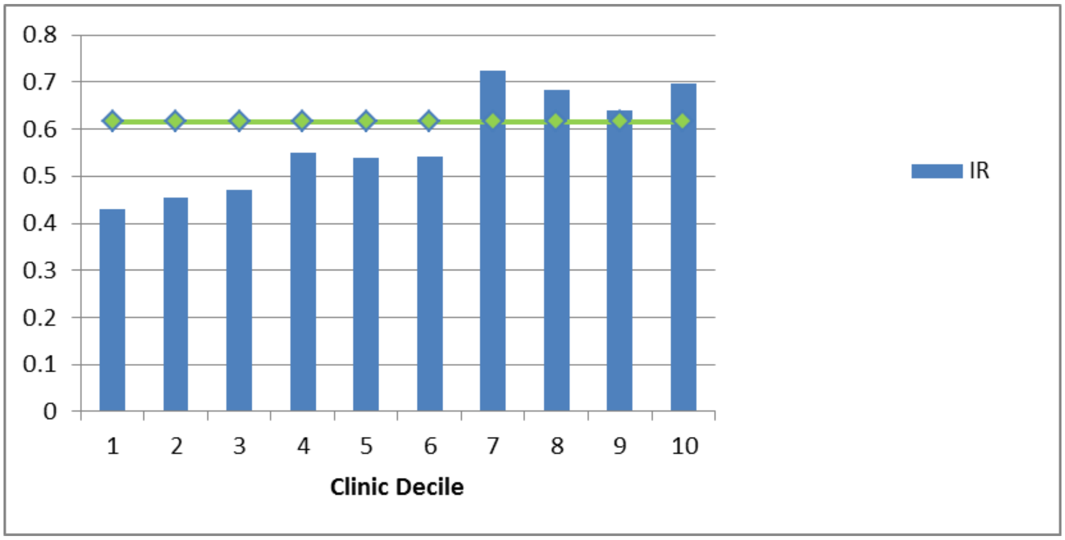


Figure 15: Performance of Good Expanded Blastocysts by Clinic Decile

Again, as shown in Figure 15, a strong upward trend is present, with a notable separation between the top 4 deciles and the remaining groups.

So, it is clear that a difference in performance exists independently in these groups. How can this be compared in the overall dataset, though? Taking a more holistic approach, consider the performance of each embryo relative to what it's score predicts in each case. This can be done by looking at the relative performance of each entry as:

$$Relative\ performance = (Embryo\ Result) - (Embyro\ Score)$$

In the example of the expanded blastocysts above, this value would be derived by the (bar height) – line. Essentially, this takes an overall weighted average of the analysis above being performed for each embryo grouping. Another way of thinking about this is: the performance of an embryo relative to what it's score would predict. Figure 16 shows the mean expected performance, measured by average embryo score at each clinic decile.

Actual performance in the chart is derived from the total implantations divided by total number of embryos transferred in that decile.

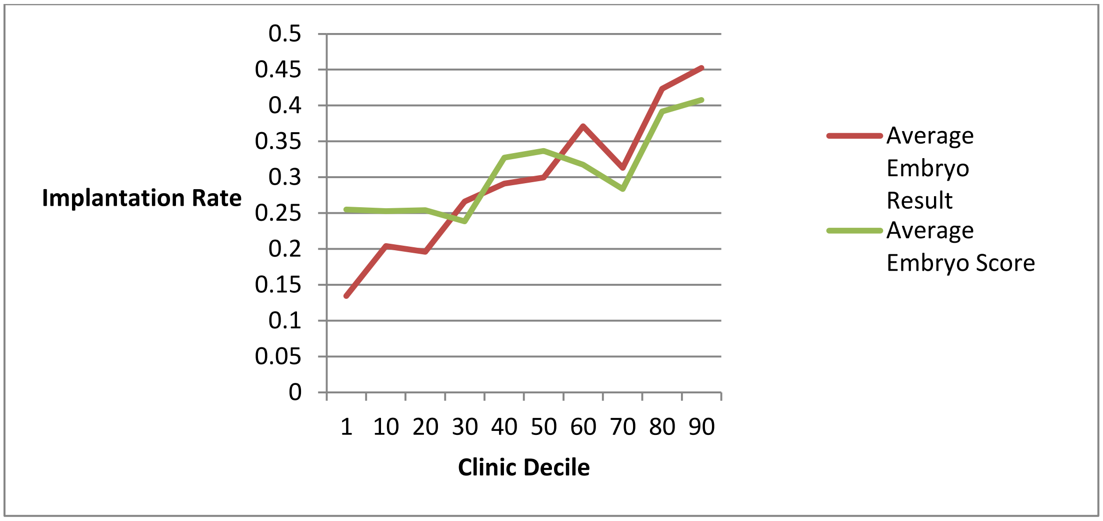


Figure 16: Implantation Rate by Clinic

Taking the difference between the two lines, mean relative performance is calculated for each decile:

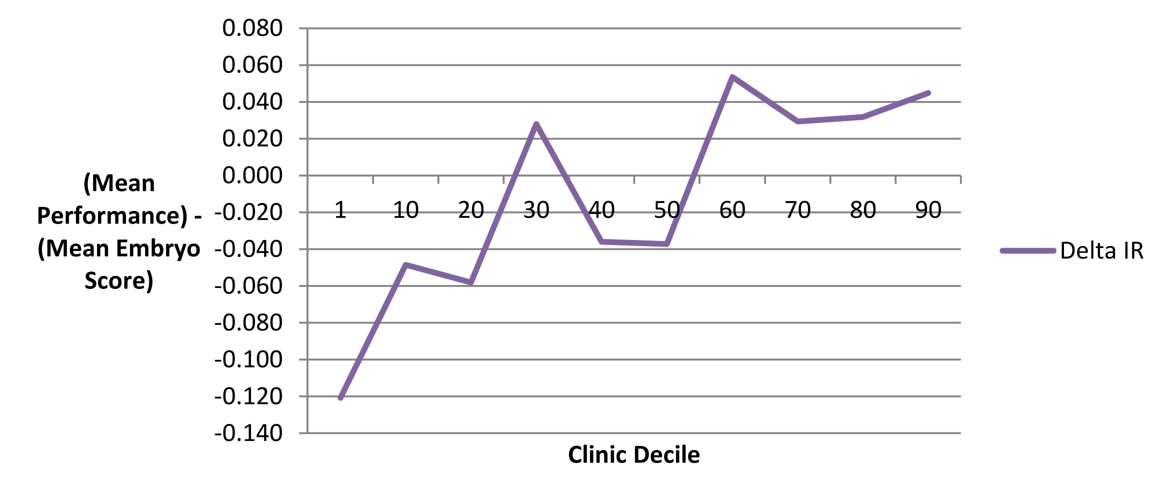


Figure 17: Mean Relative Value by Clinic

This graph is very interesting as it affirms the conclusions made within the most common embryo morphologies span the entire dataset. The spike in the 30th percentile clinics is also intriguing. Recall Figure 10 and notice that the 30th percentile clinics had the lowest overall embryo score. These clinics had particularly bad embryos morphologically and produced better than expected results with them. One possible explanation is stricter grading of embryos relative to the other practitioners.

Also, an ANOVA reveals that there is a significant difference in relative performance among the clinics (p<.001). The ANOVA results are in Table 21.

Table 21: ANOVA of Relative Performance by Clinic

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
PercentileGroup	9	49.3869	5.48744	43.1734	<.0001*
Error	23639	3004.5705	0.12710		
C. Total	23648	3053.9574			

The connecting report can be found in Appendix B. Morphologically equivalent embryos have significantly higher implantation rates at the best performing clinics.

This conclusion is not confounded by age or day of transfer and proposes opposing evidence to the argument typically made by the lower performing clinics that their lower success rates are due to patients. Granted, other factors could still be in play. Further research will be needed to examine the role of patient diagnosis.

The purpose of the clinic comparisons were to dive deeply into the differences and try to better understand why such a large disparity exists in live birth rates among clinics. The clinics included in this study are sprinkled across the rankings of live birth rate, so obviously there is a difference in live birth rates. The interesting part is breaking down the process and understanding which portions are the source of difference. It was shown that there is not a significant difference in oocyte production among clinics. Also, the translation between IR and live births per embryo transferred seemed to be relatively constant with the exception of the poorest two clinic groupings. The cause of these lowest two groupings lagging behind is unknown, although it certainly provides an opportunity for a future studies to be performed exploring causes for differences in live births per implantation.

Given the relative equality of the beginning and end of the IVF process, what remains is the portion in which the clinic is most involved: the development from oocyte to transferable embryo. The majority of the variation in live birth rate is attributed to this stage. Clinic performance in this stage was assessed and compared in this thesis.

CHAPTER 5: CONCLUSIONS

This exploratory study aimed to do several things. The initial purpose was to develop a foundation for understanding of embryo performance by applying statistical analysis techniques to a large dataset of IVF cycles. The data was thoroughly examined to see what areas can be better understood. The main accomplishment of this thesis was to propose a new embryo scoring method and to use the scores to make comparisons between clinics.

The embryo scores generated in the contingency tables are uniquely valuable because they are the first of their kind: derived directly from the average performance of that particular embryo in terms of the most relevant metric, PEIR. Implantation rate measures an embryo's ability to result in a pregnancy.

Given that the contingency tables calculate the PEIR for each treatment based on average performance of all embryos of that type, clinics now have a baseline for the relative performance of embryos of given morphologies. By comparing their historic success to the mean of the contingency tables, 30% implantation rate, they can make adjustments to their practices accordingly.

A one-dimensional embryo scoring metric was essential to facilitate the comparison made in this paper, and ultimately several conclusions drawn. First, the better clinics produce morphologically superior embryos in terms of embryo score. Additionally, it was demonstrated that embryos of identical morphology perform better at the higher percentile clinics.

Having a baseline for embryo performance allows clinics to better understand and control variation in the IVF process. Hopefully, this metric will be adopted by the community. The great thing about this metric is that the more it is used, the more reliable the mean performance estimate for each morphology combination becomes. If further studies are performed using this metric, they can juxtapose their data on top of this data to increase significance.

CHAPTER 6: RECOMMENDATIONS FOR FUTURE STUDY

The actual performance of clinics differs more than the disparity of embryo scores predict it would. This indicates that there is more to the embryo's viability than is captured by the SART descriptors. Further and more detailed embryo assessment and tracking should be considered. Additionally, experiments should be performed to understand the effects of clinic based factors. A study within a particular clinic comparing performance of embryos before and after a change is made such as temperature, air quality, or procedural method, could produce interesting results. The contingency tables generated in this study offer a basis for performing such analyses.

Also, other patient based factors could be in play. Future study exploring the exact effects on other measurable patient factors can further add to the predictability of treatment outcomes.

Another area to build upon this study is analyzing the variability in embryo quality developed from one patient. In the context of IVF, it would be very challenging to develop a paired comparison between clinics, because it is unlikely a patient will be willing to sign up for multiple IVF procedures at different clinics. However, a unique opportunity exists to compare the variability in embryo quality while removing patient effect: compare the quality of embryos from a single patient. The limitation in the SART dataset is that it only tracks the embryos that are transferred, the cream of the crop from all of the oocytes "cultivated". In order to properly perform this analysis, a morphological breakdown is necessary for all embryos grown by a clinic, not just those that are transferred.

REFERENCES

Alikani, M. , Cohen, J. , Tomkin, G. , Garrisi, G. , Mack, C. , et al. (1999). Human embryo fragmentation in vitro and its implications for pregnancy and implantation. Fertility and Sterility, 71(5), 836-842.

American Society for Reproductive Medicine. (n.d.) Embryo Morphology – SART Grading. Retrieved from http://www.asrm.org/uploadedFiles/Affiliates/SART/Members/Forms/Embryo%20Morphology.pdf

ART fact sheet. (2014, June 1). Retrieved February 17, 2015, from http://www.eshre.eu/Guidelines-and-Legal/ART-fact-sheet.aspx

Assisted Reproductive Technology (ART). (2013, November 27). Retrieved February 17, 2015, from http://www.cdc.gov/art/ART2011/

Best IVF Clinics In California For Women Under 35 Using Fresh Embryos. (2013, January 1). Retrieved March 4, 2015, from http://fertilitysuccessrates.com/report/California/women-under-35/data.html

Bhattacharya, S. , & Templeton, A. (2004). What is the most relevant standard of success in assisted reproduction? redefining success in the context of elective single embryo transfer: Evidence, intuition and financial reality. Human Reproduction (Oxford, England), 19(9), 1939-1942.

Chuang, C. , Chen, S. , Chen, C. , Chao, K. , Ho, H. , et al. (2003). Age is a better predictor of pregnancy potential than basal follicle-stimulating hormone levels in women undergoing in vitro fertilization. Fertility and Sterility, 79(1), 63-68.

Clomiphene Citrate Challenge Test (CCCT). (n.d.). Retrieved February 17, 2015, from http://www.infertilityspecialist.com/female_infertility_tests_clomid_challenge.htm

Christiano, D. (2011, January 1). Fertility Treatment Options. Retrieved March 1, 2015, from http://www.parents.com/getting-pregnant/infertility/treatments/guide-to-fertility-methods/#In Vitro Fertilization (IVF)

Cummins, J. , Breen, T. , Harrison, K. , Shaw, J. , Wilson, L. , et al. (1986). 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 : IVF, 3(5), 284-295.

Davies, M. , Wang, J. , & Norman, R. (2004). What is the most relevant standard of success in assisted reproduction? assessing the besst index for reproduction treatment. Human Reproduction (Oxford, England), 19(5), 1049-1051.

De Placido, G. , Wilding, M. , Strina, I. , Alviggi, E. , Alviggi, C. , et al. (2002). High outcome predictability after ivf using a combined score for zygote and embryo morphology and growth rate. Human Reproduction (Oxford, England), 17(9), 2402-2409.

Desai, N. , Goldstein, J. , Rowland, D. , & Goldfarb, J. (2000). Morphological evaluation of human embryos and derivation of an embryo quality scoring system specific for day 3 embryos: A preliminary study. Human Reproduction (Oxford, England), 15(10), 2190-2196.

Frazier, L. , Grainger, D. , Schieve, L. , & Toner, J. (2004). Follicle-stimulating hormone and estradiol levels independently predict the success of assisted reproductive technology treatment. Fertility and Sterility, 82(4), 834-840.

Gibbons, W. , Grainger, D. , Cedars, M. , Jain, T. , Klein, N. , et al. (2007). Continuous quality improvement and assisted reproductive technology multiple gestations: Some progress, some answers, more questions. Fertility and Sterility, 88(2), 301-304.

GIORGETTI, C. , TERRIOU, P. , AUQUIER, P. , HANS, E. , SPACH, J. , et al. (1995). Embryo score to predict implantation after in-vitro fertilization - based on 957 single embryo transfers. Human Reproduction, 10(9), 2427-2431.

Heijnen, E. , Macklon, N. , & Fauser, B. (2004). What is the most relevant standard of success in assisted reproduction? the next step to improving outcomes of ivf: Consider the whole treatment. Human Reproduction (Oxford, England), 19(9), 1936-1938.

Hu, Y. , Maxson, W. , Hoffman, D. , Ory, S. , Eager, S. , et al. (1998). Maximizing pregnancy rates and limiting higher-order multiple conceptions by determining the optimal number of embryos to transfer based on quality. Fertility and Sterility, 69(4), 650-657.

Jungheim, E. , Ryan, G. , Levens, E. , Cunningham, A. , Macones, G. , et al. (2010). Embryo transfer practices in the united states: A survey of clinics registered with the society for assisted reproductive technology. Fertility and Sterility, 94(4), 1432-1436.

Machtinger, R. , & Racowsky, C. (2013). Morphological systems of human embryo assessment and clinical evidence. Reproductive Biomedicine Online, 26(3), 210-221.

Mauer, E. (n.d.). How Much Fertility Treatments Cost. Retrieved February 17, 2015, from http://www.thebump.com/a/how-much-fertility-treatments-cost

Messinis, I. , & Domali, E. (2004). What is the most relevant standard of success in assisted reproduction? should besst really be the primary endpoint for assisted production?. Human Reproduction (Oxford, England), 19(9), 1933-1935.

Min, J. , Breheny, S. , MacLachlan, V. , & Healy, D. (2004). What is the most relevant standard of success in assisted reproduction? the singleton, term gestation, live birth rate per cycle initiated: The besst endpoint for assisted reproduction. Human Reproduction (Oxford, England), 19(1), 3-7.

Nordqvist, C. (2014, August 19). What is infertility? What causes infertility? How is infertility treated? Retrieved February 17, 2015, from http://www.medicalnewstoday.com/articles/165748.php

Patrizio, P. , & Sakkas, D. (2009). From oocyte to baby: A clinical evaluation of the Biological Efficiency of In-vitro Fertilization. Fertility and Sterility, 91(4), 1061-1066.

Pinborg, A. , Loft, A. , Ziebe, S. , & Nyboe Andersen, A. (2004). What is the most relevant standard of success in assisted reproduction? I s there a single 'parameter of excellence'?. Human Reproduction (Oxford, England), 19(5), 1052-1054.

Racowsky, C. , Combelles, C. , Nureddin, A. , Pan, Y. , Finn, A. , et al. (2003). Day 3 and day 5 morphological predictors of embryo viability. Reproductive BioMedicine Online, 6(3), 323-331.

Romito, K. (2013, November 14). In Vitro Fertilization (IVF) Risks, Success, Age, Donor Eggs, and More. Retrieved February 17, 2015, from http://www.webmd.com/infertility-and-reproduction/in-vitro-fertilization

Schieve, L. , & Reynolds, M. (2004). What is the most relevant standard of success in assisted reproduction?: Challenges in measuring and reporting success rates for assisted reproductive technology treatments: Chat is optimal?. Human Reproduction (Oxford, England), 19(4), 778-782.

Stern, J. , Goldman, M. , Hatasaka, H. , MacKenzie, T. , Surrey, E. , et al. (2009).Optimizing the number of cleavage stage embryos to transfer on day 3 in women 38 years of age and older: A society for assisted reproductive technology databasestudy. Fertility and Sterility, 91(3), 767-776.

Van Royen, E. , Mangelschots, K. , De Neubourg, D. , Valkenburg, M. , Van de Meerssche, M. , et al. (1999). Characterization of a top quality embryo, a step towards single-embryo transfer. Human Reproduction (Oxford, England), 14(9), 2345-2349.

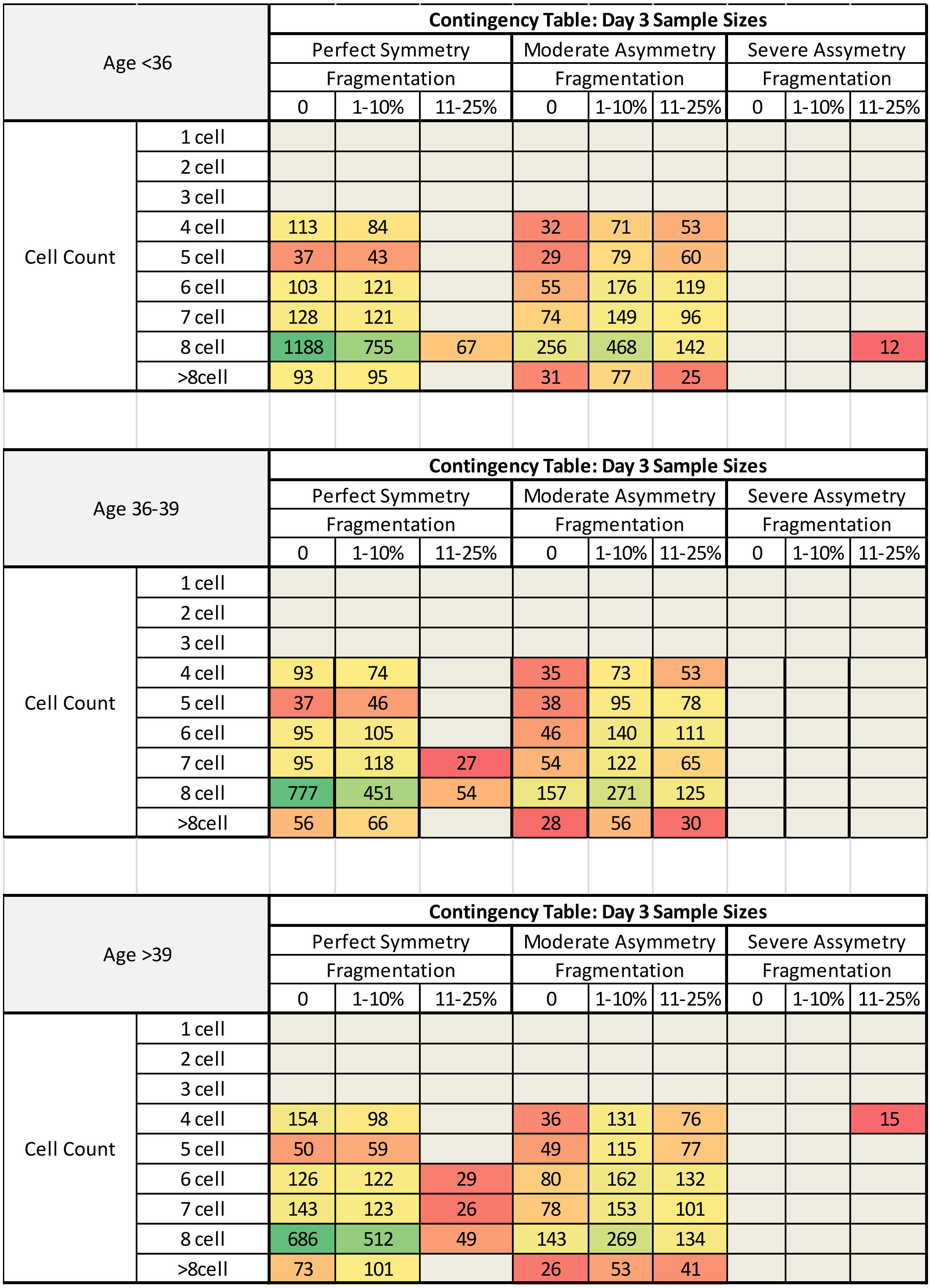
What Causes Female Infertility? (n.d.). Retrieved February 17, 2015, from https://web.stanford.edu/class/siw198q/websites/reprotech/New Ways of Making Babies/Causefem.htm

ZIEBE, S. , LOFT, A. , J.H, P. , A-G, A. , LINDENBERG, S. , et al. (2001). Embryo quality and developmental potential is compromised by age. Acta Obstetricia Et Gynecologica Scandinavica, 80(2), 169-174.

Ziebe, S. , Petersen, K. , Lindenberg, S. , Andersen, A. , Gabrielsen, A. , et al. (1997). Embryo morphology or cleavage stage: How to select the best embryos for transfer after in-vitro fertilization. Human Reproduction (Oxford, England), 12(7), 1545-1549.

APPENDICES

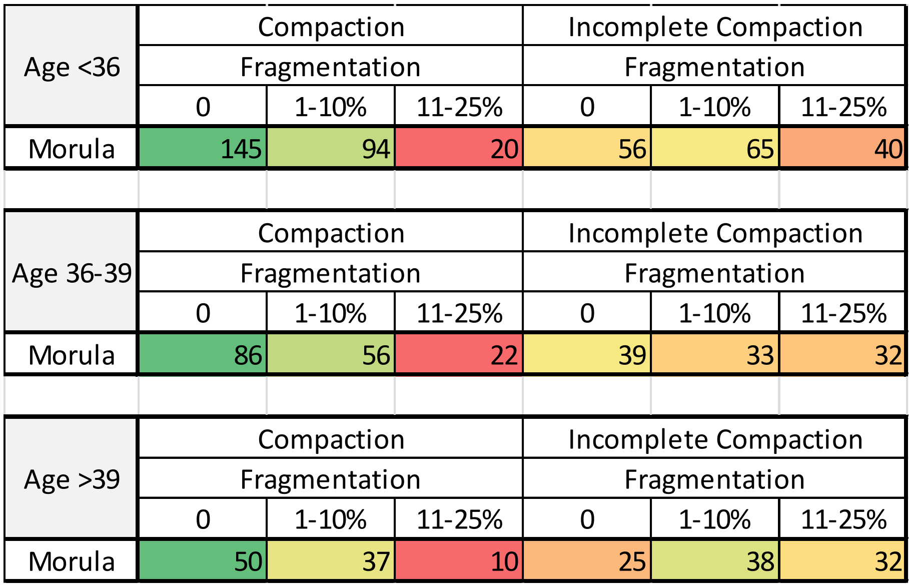
Appendix A: Contingency Table Sample Sizes



Age <36		Contingency Table: Day 3 Sample Sizes								
		Perfect Symmetry			Moderate Asymmetry			Severe Assymetry		
		Fragmentation			Fragmentation			Fragmentation		
		0	1-10%	11-25%	0	1-10%	11-25%	0	1-10%	11-25%
Cell Count	1 cell									
	2 cell									
	3 cell									
	4 cell	113	84		32	71	53			
	5 cell	37	43		29	79	60			
	6 cell	103	121		55	176	119			
	7 cell	128	121		74	149	96			
	8 cell	1188	755	67	256	468	142			12
	>8cell	93	95		31	77	25			

Age 36-39		Contingency Table: Day 3 Sample Sizes								
		Perfect Symmetry			Moderate Asymmetry			Severe Assymetry		
		Fragmentation			Fragmentation			Fragmentation		
		0	1-10%	11-25%	0	1-10%	11-25%	0	1-10%	11-25%
Cell Count	1 cell									
	2 cell									
	3 cell									
	4 cell	93	74		35	73	53			
	5 cell	37	46		38	95	78			
	6 cell	95	105		46	140	111			
	7 cell	95	118	27	54	122	65			
	8 cell	777	451	54	157	271	125			
	>8cell	56	66		28	56	30			

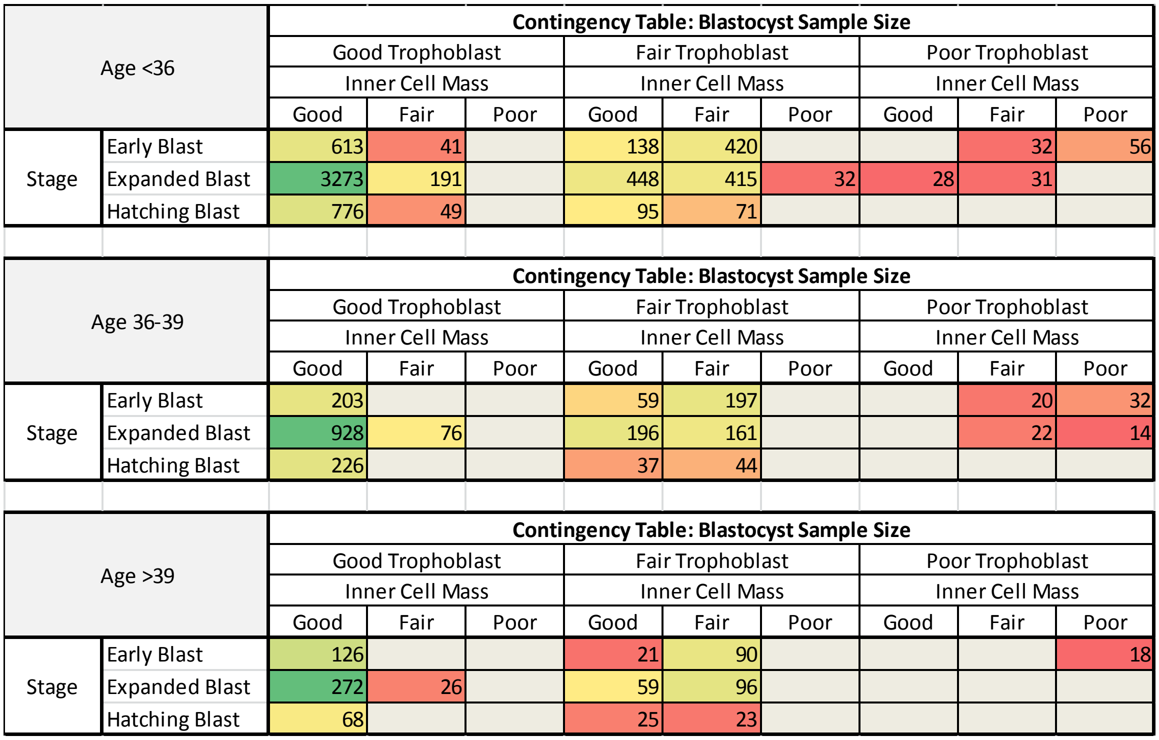
Age >39		Contingency Table: Day 3 Sample Sizes								
		Perfect Symmetry			Moderate Asymmetry			Severe Assymetry		
		Fragmentation			Fragmentation			Fragmentation		
		0	1-10%	11-25%	0	1-10%	11-25%	0	1-10%	11-25%
Cell Count	1 cell									
	2 cell									
	3 cell									
	4 cell	154	98		36	131	76			15
	5 cell	50	59		49	115	77			
	6 cell	126	122	29	80	162	132			
	7 cell	143	123	26	78	153	101			
	8 cell	686	512	49	143	269	134			
	>8cell	73	101		26	53	41			



Age <36	Compaction			Incomplete Compaction		
	Fragmentation			Fragmentation		
	0	1-10%	11-25%	0	1-10%	11-25%
Morula	145	94	20	56	65	40

Age 36-39	Compaction			Incomplete Compaction		
	Fragmentation			Fragmentation		
	0	1-10%	11-25%	0	1-10%	11-25%
Morula	86	56	22	39	33	32

Age >39	Compaction			Incomplete Compaction		
	Fragmentation			Fragmentation		
	0	1-10%	11-25%	0	1-10%	11-25%
Morula	50	37	10	25	38	32



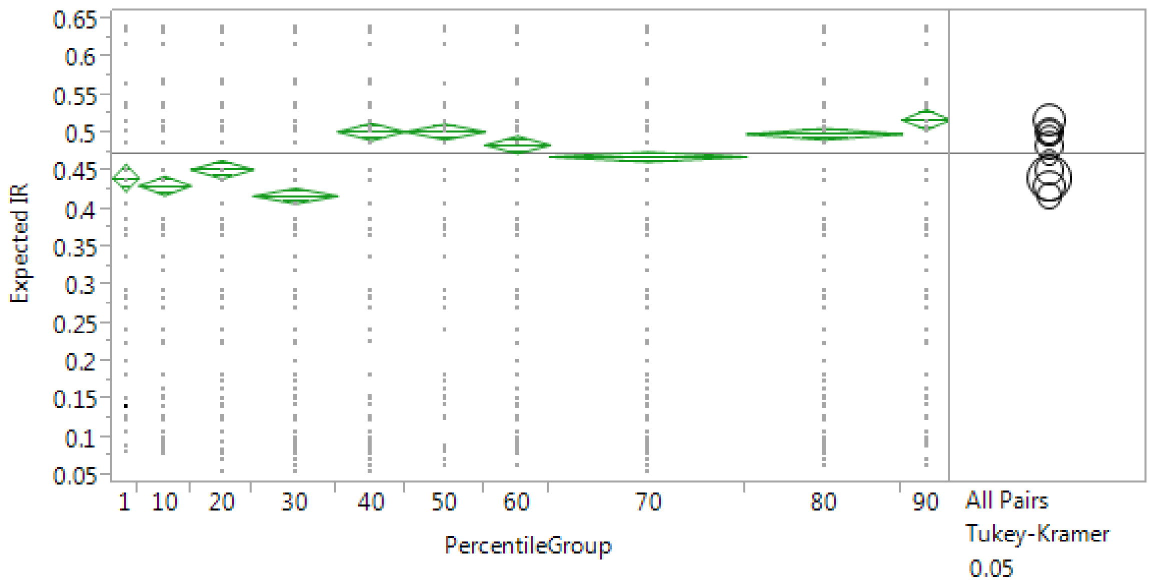
Age <36		Contingency Table: Blastocyst Sample Size								
		Good Trophoblast			Fair Trophoblast			Poor Trophoblast		
		Inner Cell Mass			Inner Cell Mass			Inner Cell Mass		
		Good	Fair	Poor	Good	Fair	Poor	Good	Fair	Poor
Stage	Early Blast	613	41		138	420			32	56
	Expanded Blast	3273	191		448	415	32	28	31	
	Hatching Blast	776	49		95	71				

Age 36-39		Contingency Table: Blastocyst Sample Size								
		Good Trophoblast			Fair Trophoblast			Poor Trophoblast		
		Inner Cell Mass			Inner Cell Mass			Inner Cell Mass		
		Good	Fair	Poor	Good	Fair	Poor	Good	Fair	Poor
Stage	Early Blast	203			59	197			20	32
	Expanded Blast	928	76		196	161			22	14
	Hatching Blast	226			37	44				

Age >39		Contingency Table: Blastocyst Sample Size								
		Good Trophoblast			Fair Trophoblast			Poor Trophoblast		
		Inner Cell Mass			Inner Cell Mass			Inner Cell Mass		
		Good	Fair	Poor	Good	Fair	Poor	Good	Fair	Poor
Stage	Early Blast	126			21	90				18
	Expanded Blast	272	26		59	96				
	Hatching Blast	68			25	23				

Appendix B: JMP Outputs

B.1 Comparison of Day 5 scores by clinic



Connecting Letters Report

```
Level                                    Mean
90        A                        0.51803801
40        A   B                    0.50224059
50        A   B                    0.50210710
80        A   B                    0.49916117
60            B   C                0.48506333
70                C   D            0.46930236
20                    D   E        0.45253568
1                     D   E   F    0.44165333
10                        E   F    0.43135228
30                            F    0.41796677
```

Levels not connected by same letter are significantly different.

B.2 Relative Value

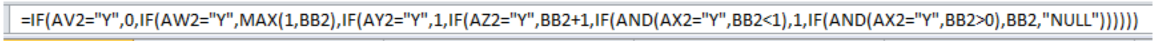
Connecting Letters Report

Level				**Mean**
60	A			0.0535446
90	A			0.0447630
80	A			0.0317730
70	A			0.0294183
30	A			0.0279788
40		B		-0.0360669
50		B		-0.0372175
10		B		-0.0485057
20		B		-0.0580537
1			C	-0.1208109

Levels not connected by same letter are significantly different.

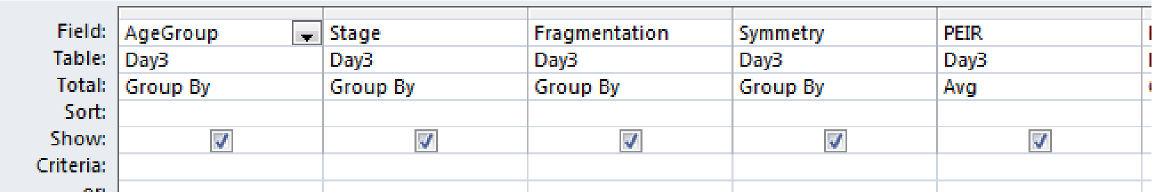
Appendix C: Formulas and Queries

C.1 Implantations Calculation



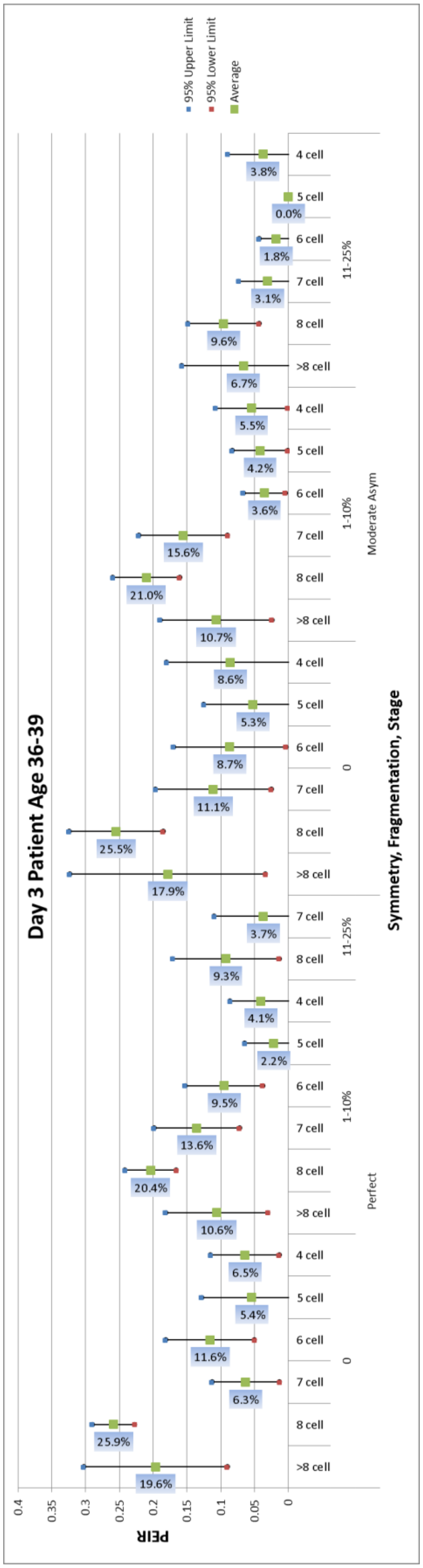
```
=IF(AV2="Y",0,IF(AW2="Y",MAX(1,BB2),IF(AY2="Y",1,IF(AZ2="Y",BB2+1,IF(AND(AX2="Y",BB2<1),1,IF(AND(AX2="Y",BB2>0),BB2,"NULL"))))))
```

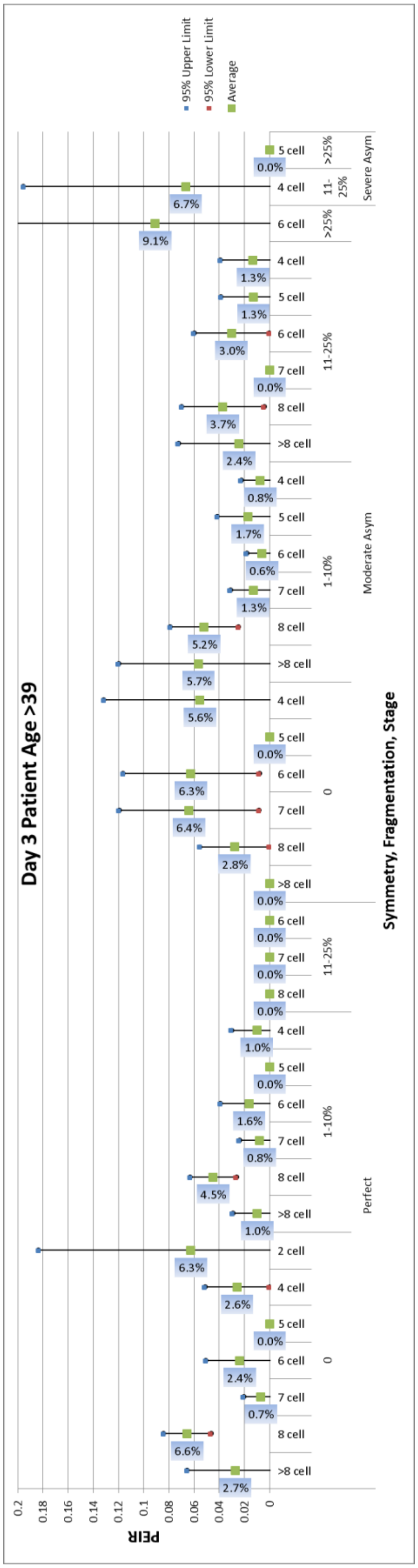
C.2 Contingency Generation Query

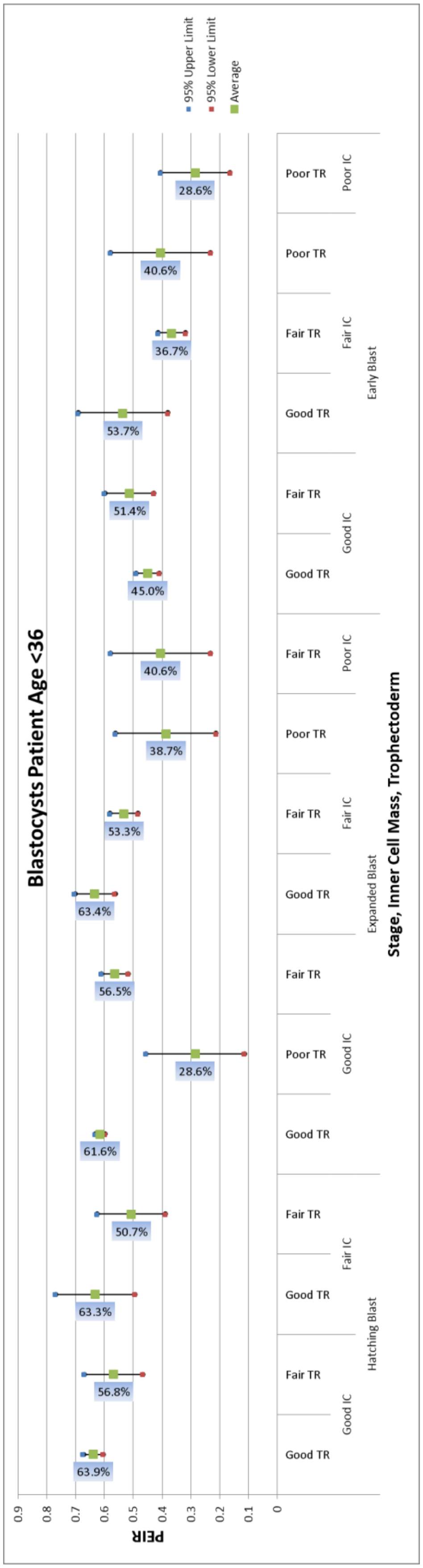


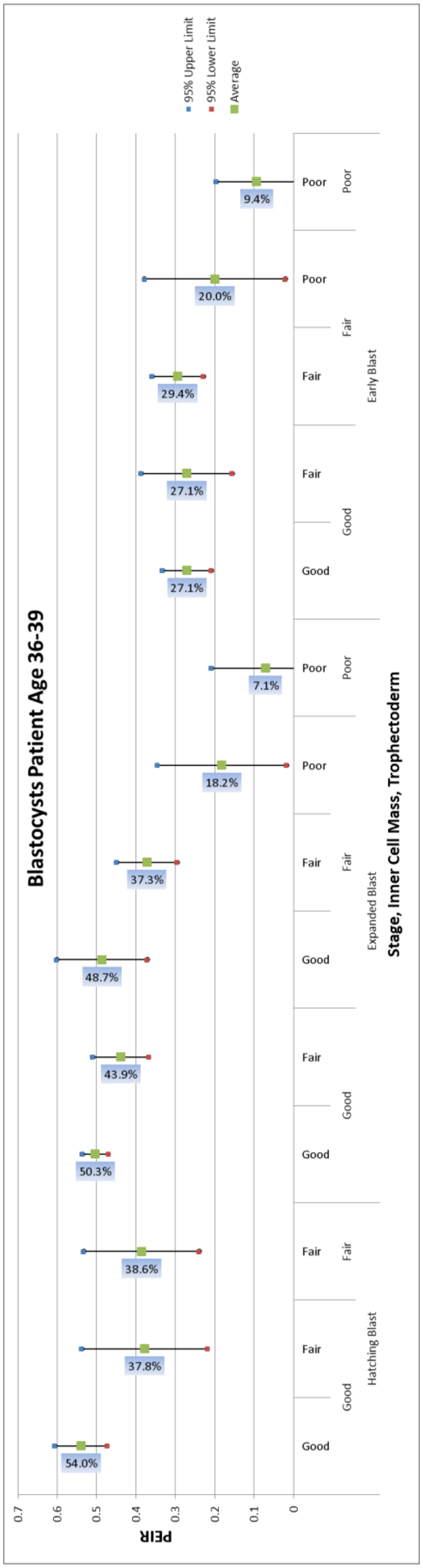
Field:	AgeGroup	Stage	Fragmentation	Symmetry	PEIR
Table:	Day3	Day3	Day3	Day3	Day3
Total:	Group By	Group By	Group By	Group By	Avg
Sort:					
Show:	☑	☑	☑	☑	☑
Criteria:					

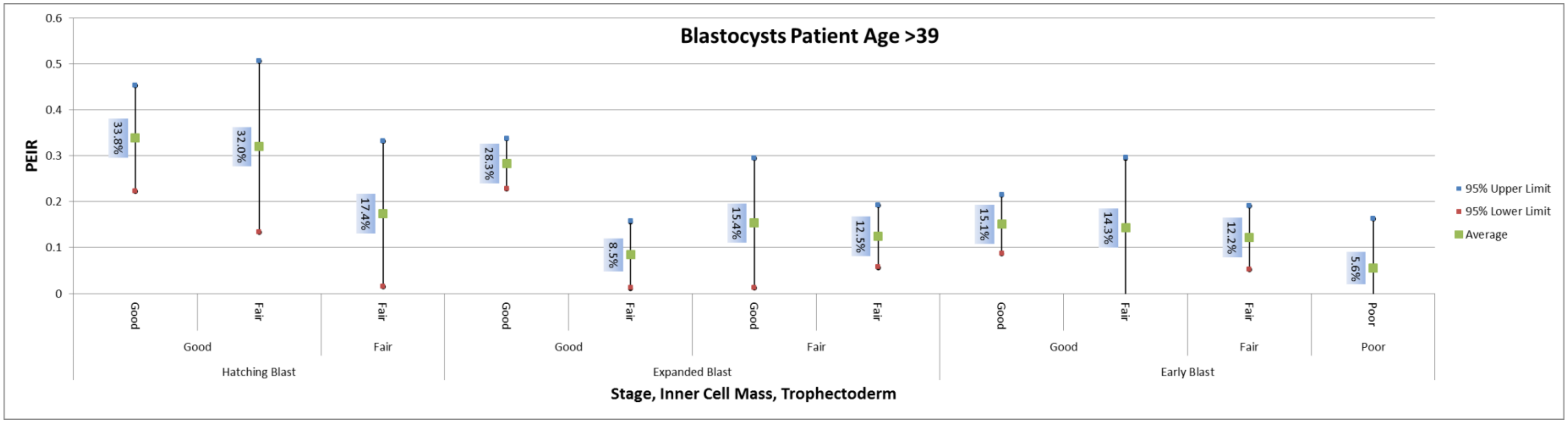
Appendix D: Contingency Table Confidence Ranges


Day 3 Patient Age 36-39
PEIR
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Symmetry, Fragmentation, Stage
■ 95% Upper Limit
■ 95% Lower Limit
■ Average
Perfect
0
>8 cell 19.6%
8 cell 25.9%
7 cell 6.3%
6 cell 11.6%
5 cell 5.4%
4 cell 6.5%
1-10%
>8 cell 10.6%
8 cell 20.4%
7 cell 13.6%
6 cell 9.5%
5 cell 2.2%
4 cell 4.1%
11-25%
8 cell 9.3%
7 cell 3.7%
Moderate Asym
0
>8 cell 17.9%
8 cell 25.5%
7 cell 11.1%
6 cell 8.7%
5 cell 5.3%
4 cell 8.6%
1-10%
>8 cell 10.7%
8 cell 21.0%
7 cell 15.6%
6 cell 3.6%
5 cell 4.2%
4 cell 5.5%
11-25%
>8 cell 6.7%
8 cell 9.6%
7 cell 3.1%
6 cell 1.8%
5 cell 0.0%
4 cell 3.8%


Day 3 Patient Age >39
PEIR
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Symmetry, Fragmentation, Stage
■ 95% Upper Limit
■ 95% Lower Limit
■ Average
Perfect
0
>8 cell 2.7%
8 cell 6.6%
7 cell 0.7%
6 cell 2.4%
5 cell 0.0%
4 cell 2.6%
2 cell 6.3%
1-10%
>8 cell 1.0%
8 cell 4.5%
7 cell 0.8%
6 cell 1.6%
5 cell 0.0%
4 cell 1.0%
11-25%
8 cell 0.0%
7 cell 0.0%
6 cell 0.0%
Moderate Asym
0
>8 cell 0.0%
8 cell 2.8%
7 cell 6.4%
6 cell 6.3%
5 cell 0.0%
4 cell 5.6%
1-10%
>8 cell 5.7%
8 cell 5.2%
7 cell 1.3%
6 cell 0.6%
5 cell 1.7%
4 cell 0.8%
11-25%
>8 cell 2.4%
8 cell 3.7%
7 cell 0.0%
6 cell 3.0%
5 cell 1.3%
4 cell 1.3%
Severe Asym
>25%
6 cell 9.1%
11-25%
4 cell 6.7%
>25%
5 cell 0.0%


Blastocysts Patient Age <36
PEIR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
95% Upper Limit
95% Lower Limit
Average
63.9%
56.8%
63.3%
50.7%
61.6%
28.6%
56.5%
63.4%
53.3%
38.7%
40.6%
45.0%
51.4%
53.7%
36.7%
40.6%
28.6%
Good TR
Fair TR
Good TR
Fair TR
Good TR
Poor TR
Fair TR
Good TR
Fair TR
Poor TR
Fair TR
Good TR
Fair TR
Good TR
Fair TR
Poor TR
Poor TR
Good IC
Fair IC
Good IC
Fair IC
Poor IC
Good IC
Fair IC
Poor IC
Hatching Blast
Expanded Blast
Early Blast
Stage, Inner Cell Mass, Trophectoderm


Blastocysts Patient Age 36-39
PEIR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
95% Upper Limit
95% Lower Limit
Average
54.0%
37.8%
38.6%
50.3%
43.9%
48.7%
37.3%
18.2%
7.1%
27.1%
27.1%
29.4%
20.0%
9.4%
Good
Fair
Fair
Good
Fair
Good
Fair
Poor
Poor
Good
Fair
Fair
Poor
Poor
Good
Fair
Good
Fair
Poor
Good
Fair
Poor
Hatching Blast
Expanded Blast
Early Blast
Stage, Inner Cell Mass, Trophectoderm


Blastocysts Patient Age >39
PEIR
0.6
0.5
0.4
0.3
0.2
0.1
0
33.8%
32.0%
17.4%
28.3%
8.5%
15.4%
12.5%
15.1%
14.3%
12.2%
5.6%
Good
Fair
Fair
Good
Fair
Good
Fair
Good
Fair
Fair
Poor
Good
Fair
Good
Fair
Good
Fair
Poor
Hatching Blast
Expanded Blast
Early Blast
Stage, Inner Cell Mass, Trophectoderm
95% Upper Limit
95% Lower Limit
Average