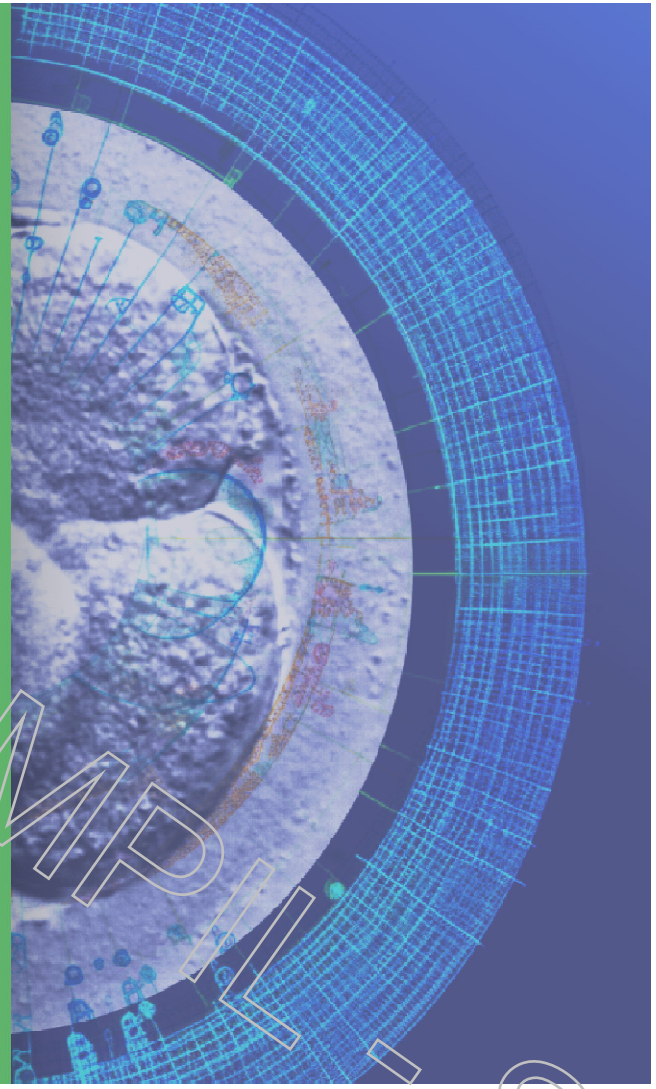
The development of an Artificial Intelligence screening test for embryo evaluation: XAI, ploidy and ethical considerations

Daniel S. Seidman, MD, MMSc

Associate Professor

Faculty of Medical & Health Sciences
Tel Aviv University



DISCLOSURE

Co-founder & Chief Medical Officer

The Decline of Embryology in IVF?

- In recent years it seems that embryology is losing grounds and instead of looking at the embryos and improving our skills, embryos are left in the incubators for 5 days and no real embryo analysis is taking place.
- AI, and the creation of a digital embryology, is bound **not to replace the embryologist but to bring back the importance of the embryology science in the lab**.
- When we established AIVF one of the goals was to improve our understanding of embryology and our ability to better understand embryo development.

AIVF's Vision: Empowering Embryology with AI

- I am proud to say that we are now showing that AI can augment our capabilities, and embryologists can better understand the science through AI.
- We have long been interested in proving that AI can discover new features, and we now know it is a reality.
- What I would like to present here is our recent paper, published in Nature, that shows exactly this- an interoperability of an AI model to enhance embryology science and our understanding of embryo development.

AIVF™

Ben-Gurion University of the Negev

Lab of Computational Cell Dynamics

nature communications

Article

https://doi.org/10.1038/s41467-024-51136-9

Visual interpretability of image-based classification models by generative latent space disentanglement applied to in vitro fertilization

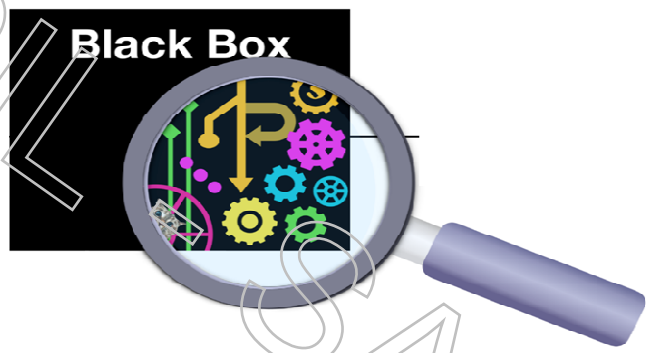
Received: 7 November 2023

Accepted: 31 July 2024

Oded Rotem[1], Tamar Schwartz[2], Ron Maor[2], Yishay Tauber[2], Maya Tsarfati Shapiro[2], Marcos Meseguer[3,4], Daniella Gilboa[2], Daniel S. Seidman[2,5] & Assaf Zaritsky[1]

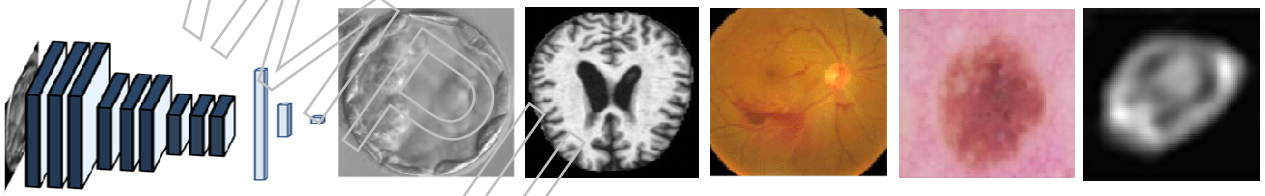
This talk is about

How can we better understand what AI models are learning



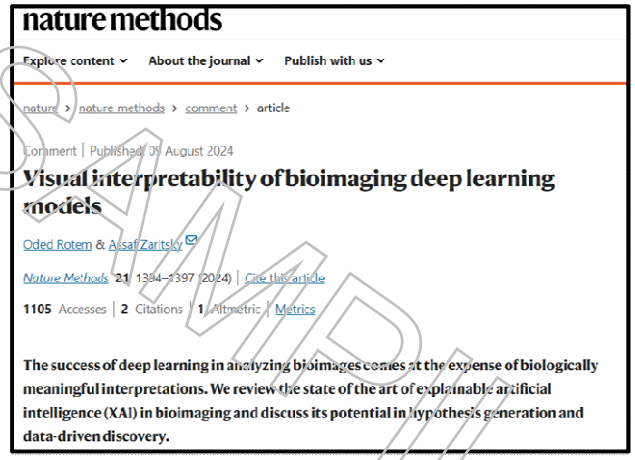
- Interpretability - deciphering the AI model
- IVF overview and the use of AI
- Our interpretability methodology
- Application in IVF

The need for better biomedical-imaging interpretability



Increase trust and reliability

- Give a reason for the diagnosis
- Intuitive explanation
- Insight to biological processes



nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › comment › article

Comment | Published: 09 August 2024

Visual interpretability of bioimaging deep learning models

Oded Rotem & Assaf Zaritsky

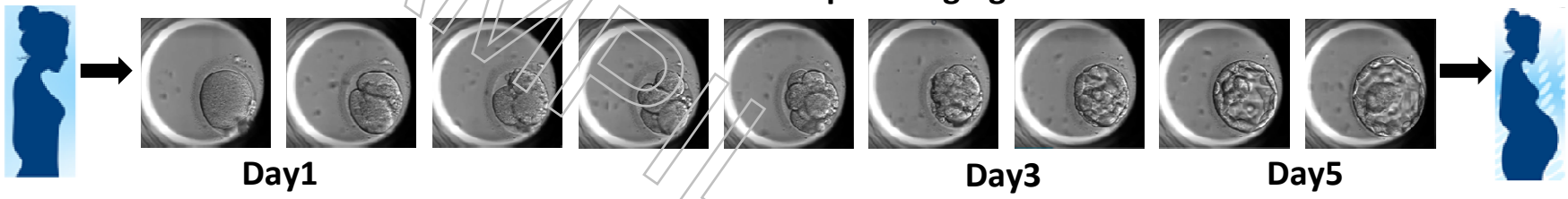
Nature Methods 21, 1394–1397 (2024) | Cite this article

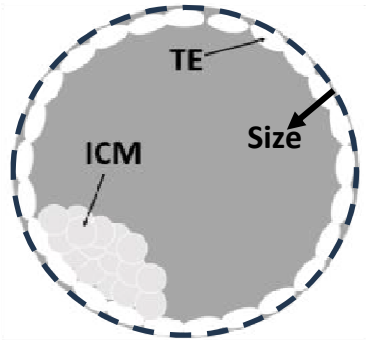
1105 Accesses | 2 Citations | 1 Altmetric | Metrics

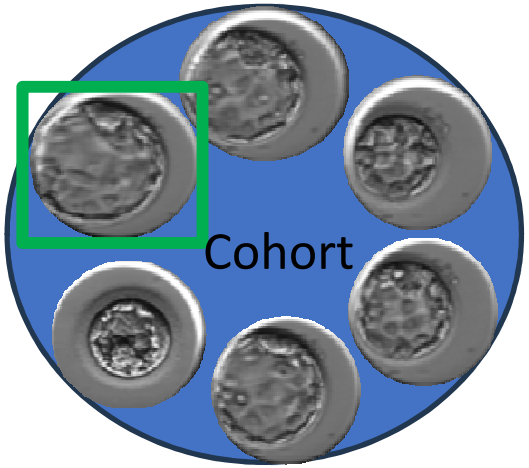
The success of deep learning in analyzing bioimages comes at the expense of biologically meaningful interpretations. We review the state of the art of explainable artificial intelligence (XAI) in bioimaging and discuss its potential in hypothesis generation and data-driven discovery.

IVF

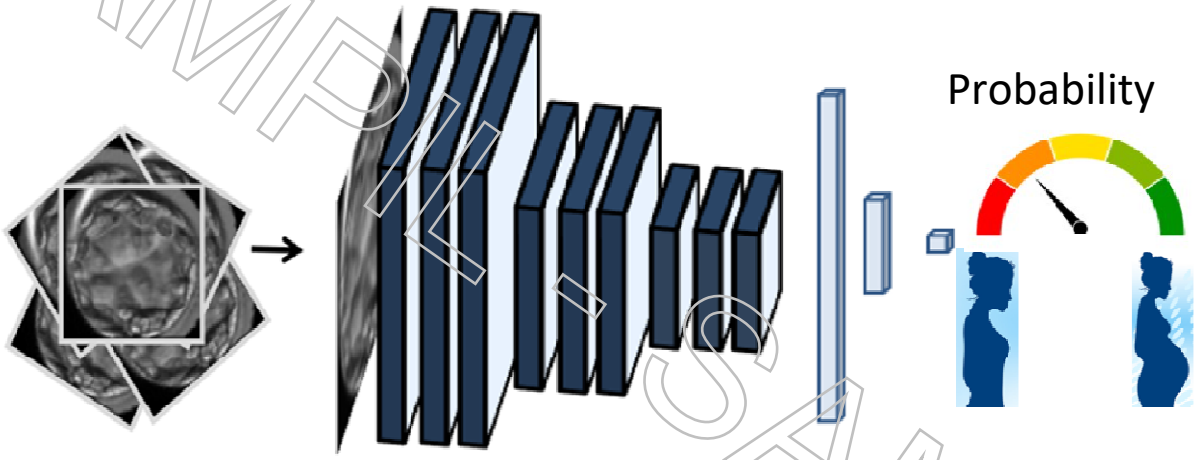
Time lapse imaging







AI assisting embryo selection

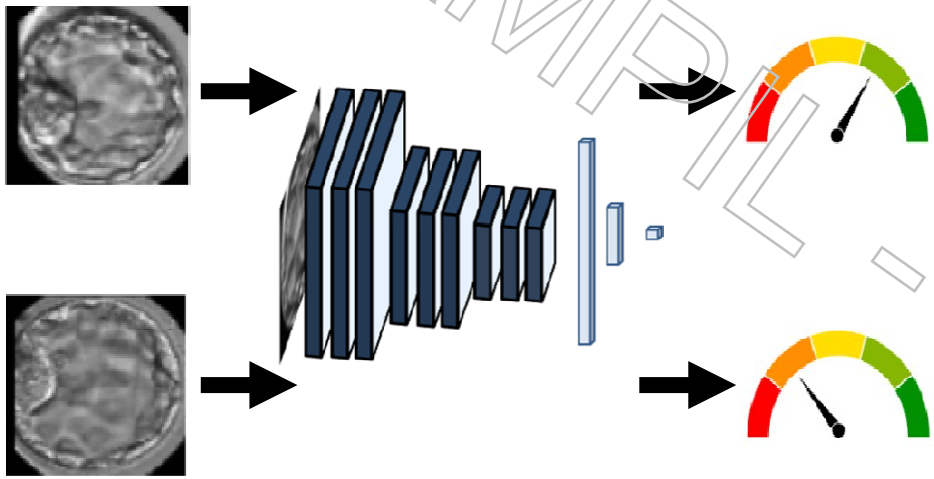


Labels:

- Grading
- Implantation
- Genetic testing

AI based classifiers have shown comparable or better results in embryo selection

Why is interpretability needed



Should I trust it?

What is it looking at?

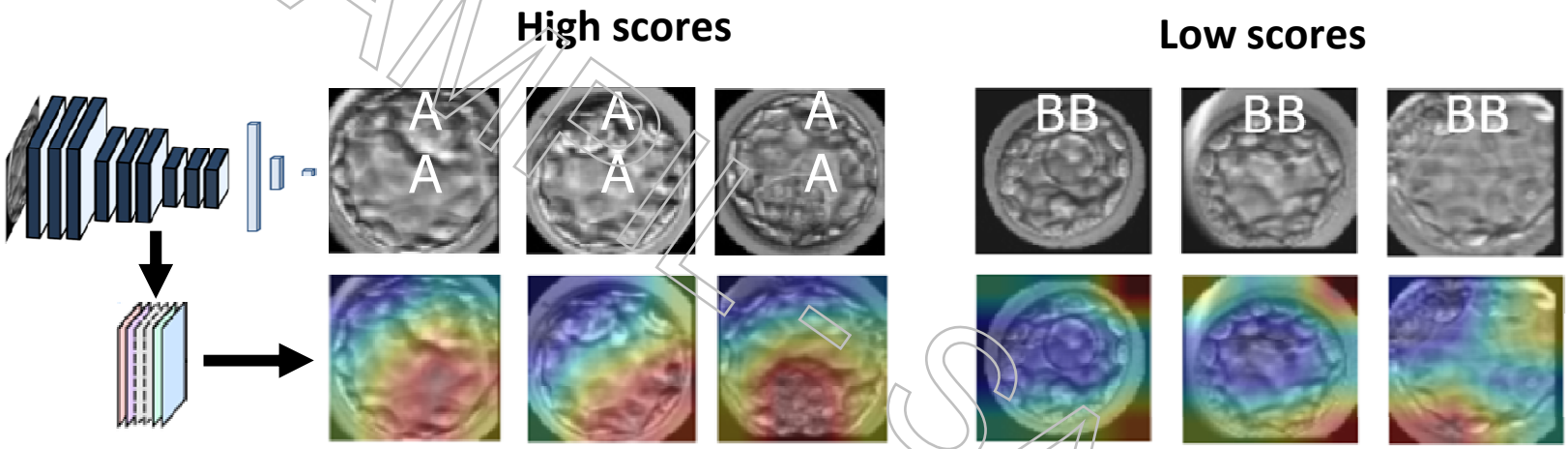
Do I agree?

Did it learn something which is unknown to us?

Is it picking up on a bias

Interpretability = explainability = XAI

Heatmaps are not enough



<u>Current methods Limitations:</u>

- Entanglement of multiple properties
- Non-local properties (size, shape, color etc.)
- Property significance

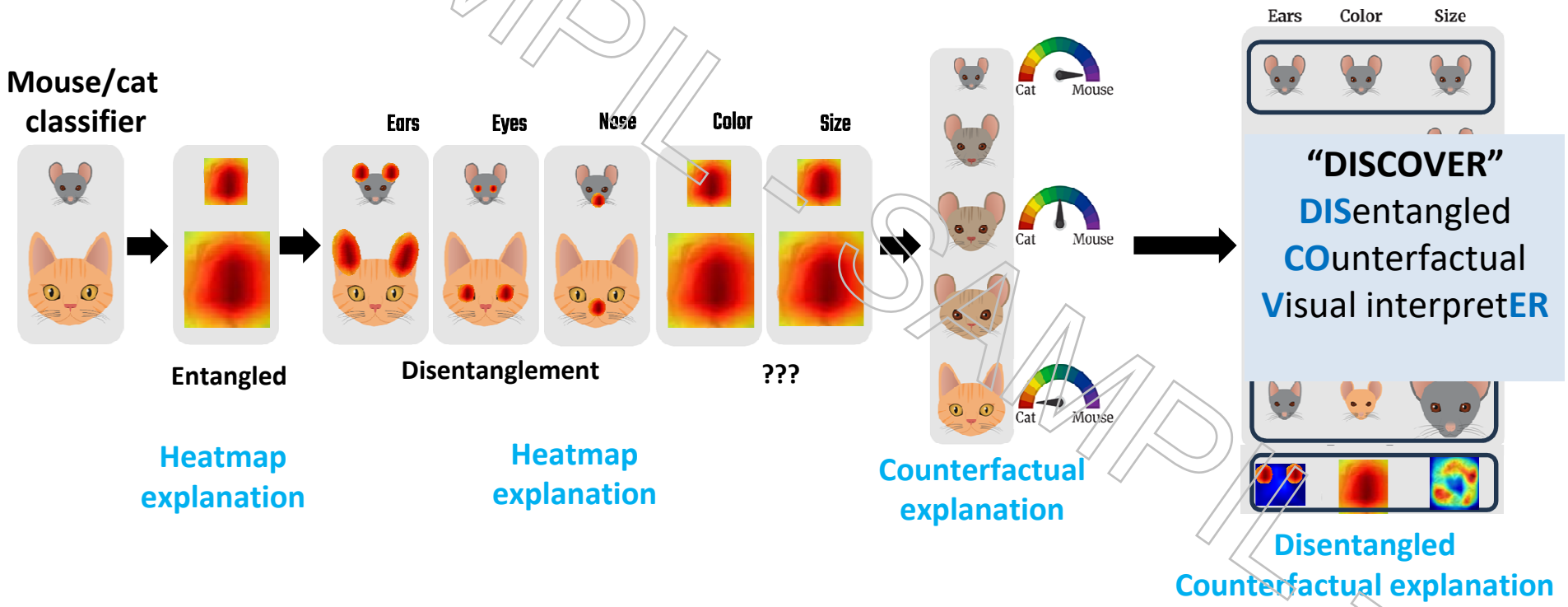
What is entanglement

- Multiple features or variables are intertwined in a model, making it difficult to separate or understand the individual influence of each feature on the final prediction.
- In the context of embryo analysis visual properties like size, shape, and cell density may all be entangled, meaning the model doesn't clearly differentiate how each feature individually impacts embryo quality. This lack of clarity makes interpreting the model's decisions challenging.
- **Disentanglement** is the process of separating these features so that each one represents a distinct property, allowing for clearer interpretation and understanding of how specific features contribute to a decision.

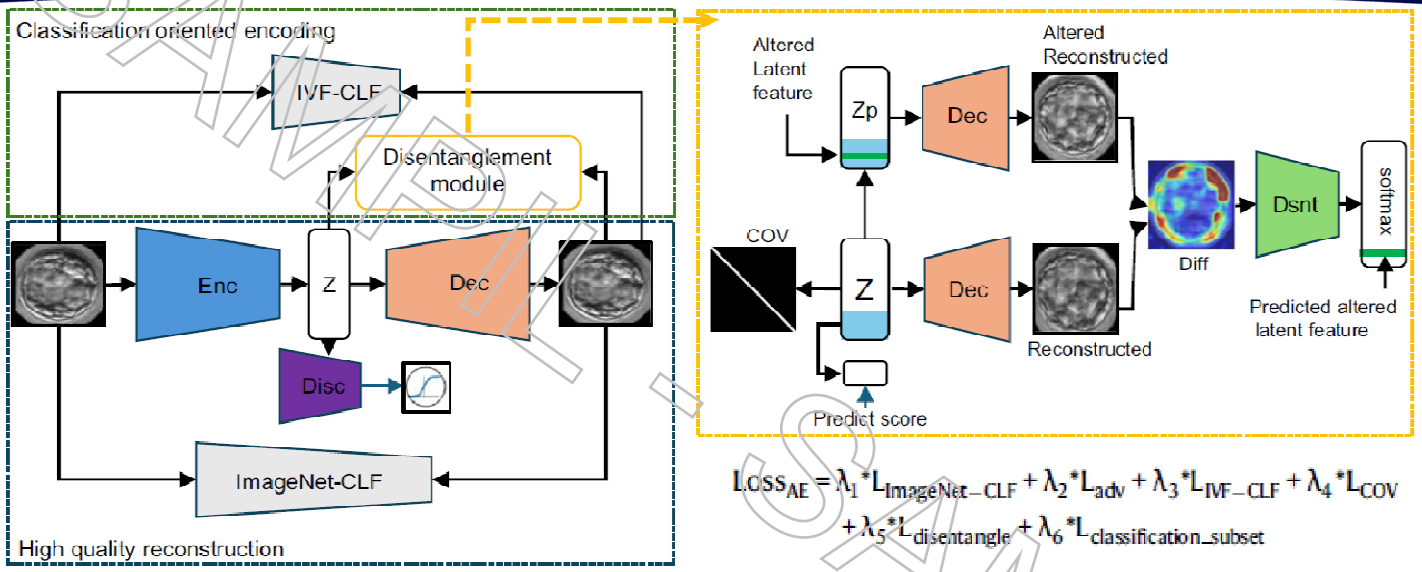
What is counterfactual explanation

- A method where small, controlled changes are made to an input (such as an image or data point) to show how the output (prediction) would change as a result.
- It answers "what-if" questions, like "What would happen if the embryo were slightly larger?"

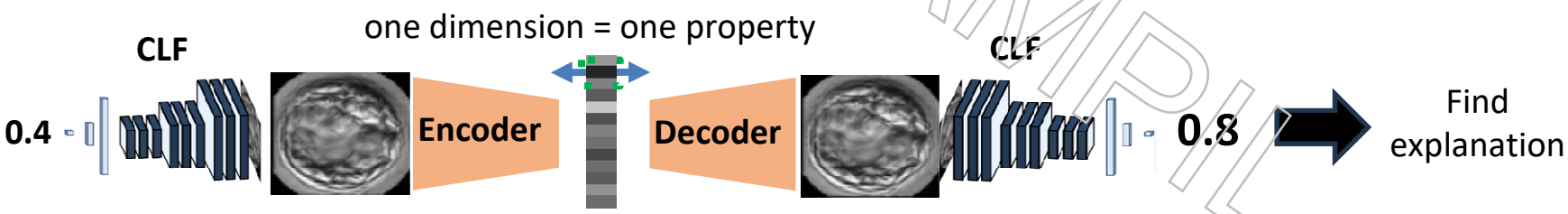
Intuitive demonstration of our method



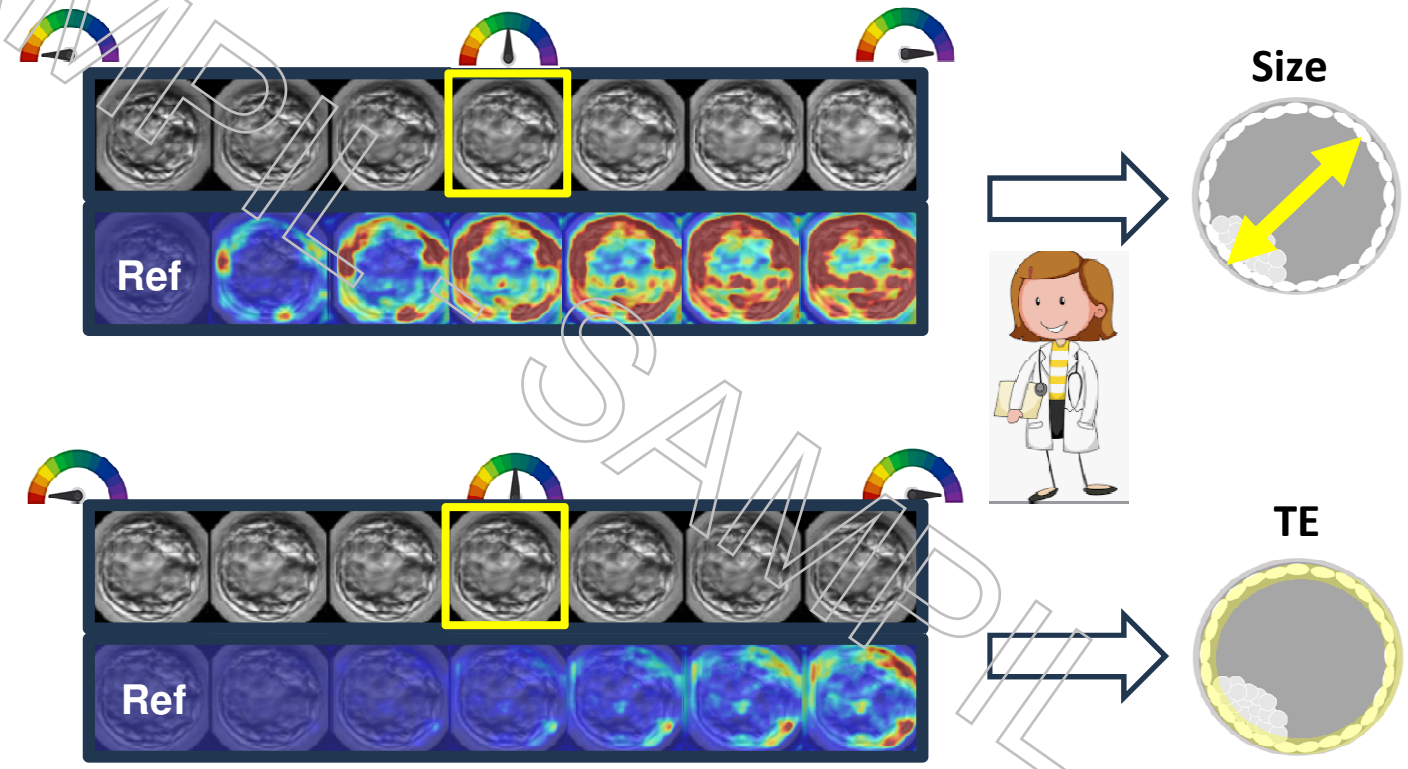
DISCOVER architecture overview



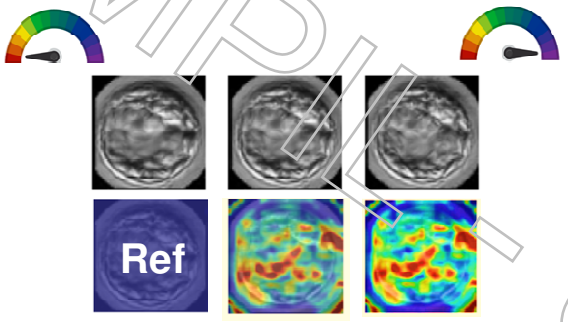
$$Loss_{AE} = \lambda_1 * L_{ImageNet-CLF} + \lambda_2 * L_{adv} + \lambda_3 * L_{IVF-CLF} + \lambda_4 * L_{COV} + \lambda_5 * L_{disentangle} + \lambda_6 * L_{classification\_subset}$$



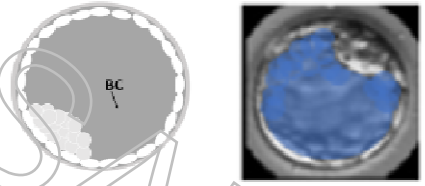
Interpreting known IVF morphology



Discovery of a new visual property

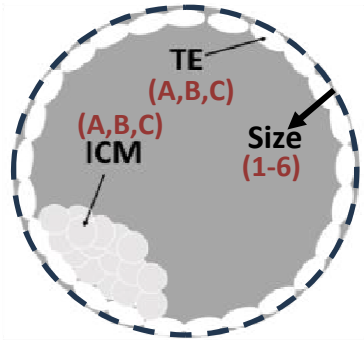


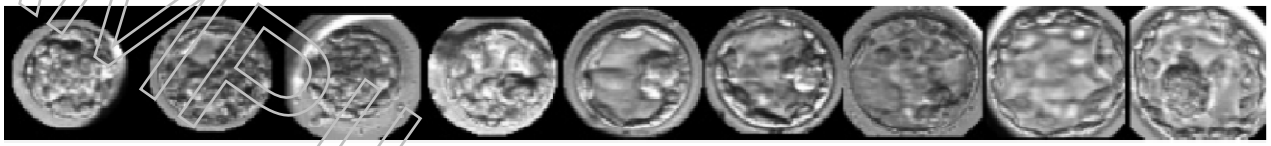
Blastocoel density: inner cavity of the blastocyst holding nutrients for the Inner Cell Mass (ICM)



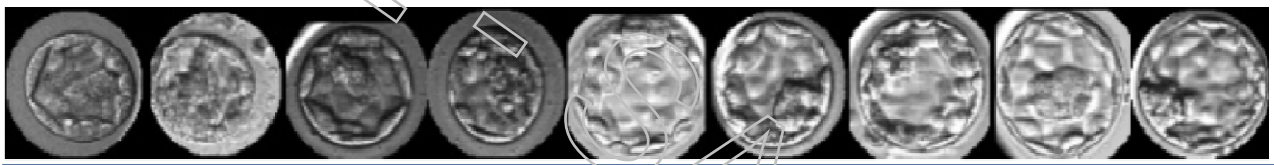
Moving from discrete to continuous grading

Grade morphologies [A,B,C] ➔ [0-1]

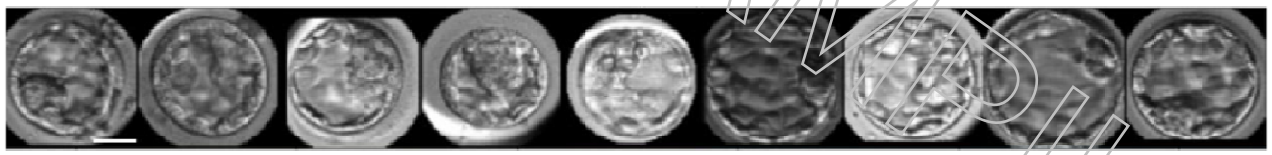




Sorting images by size



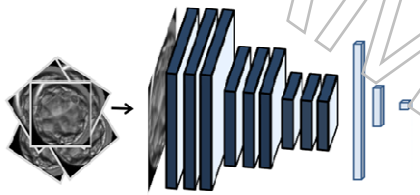
Sorting images by trophectoderm



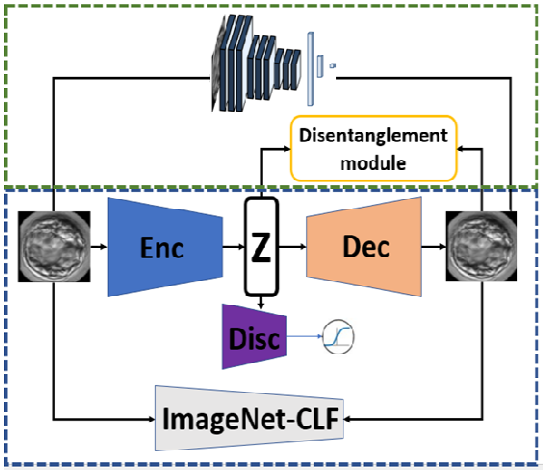
Sorting images by Blastocoel

Full interpretability process

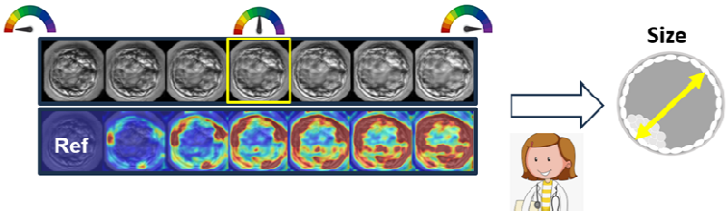
Step 0: Classifier training



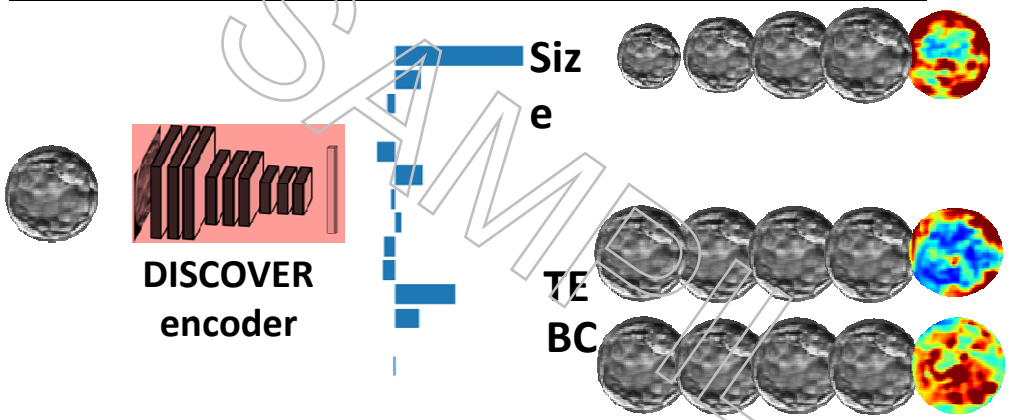
Step 1: DISCOVER training



Step 2: Global interpretability



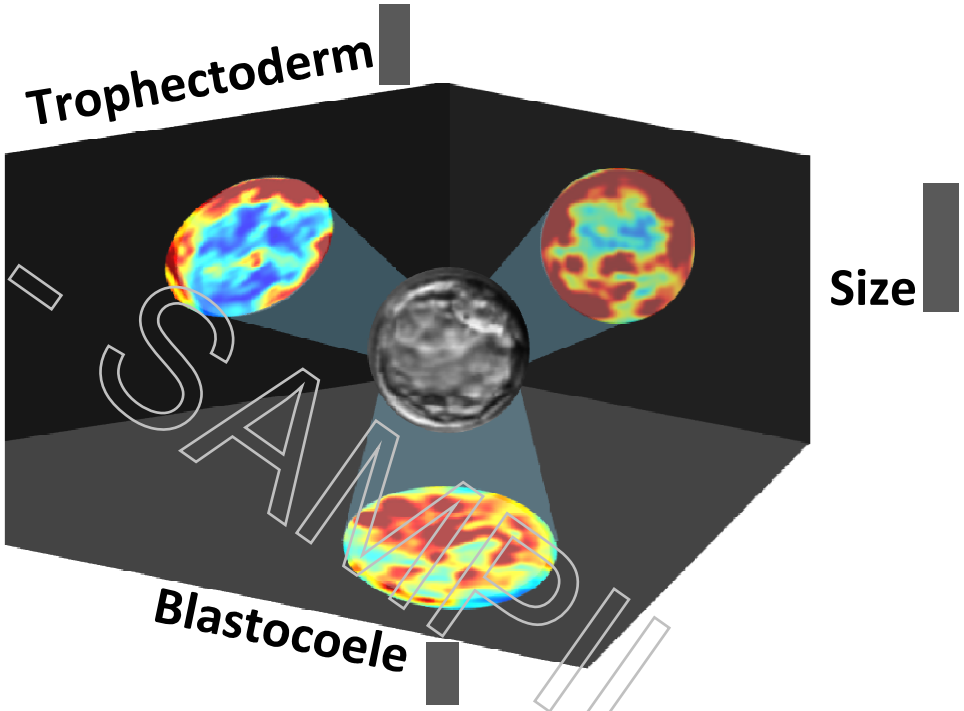
Step 3: Real-time instance interpretability



Summary

DISCOVER Capabilities:

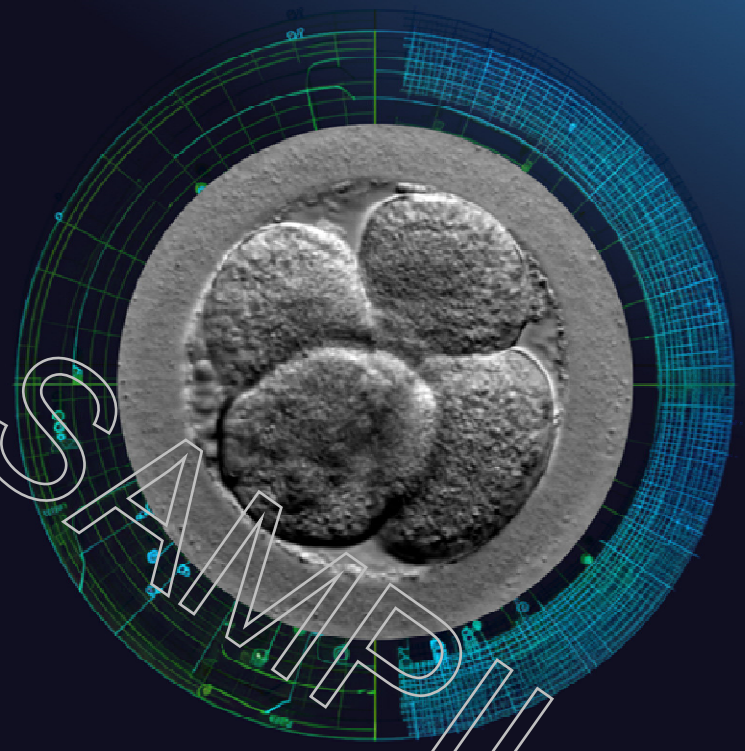
- Disentangled properties
- Intuitive Interpretability
- Discovery of new properties
- Global & instance interpretability
- Property significance



Introducing

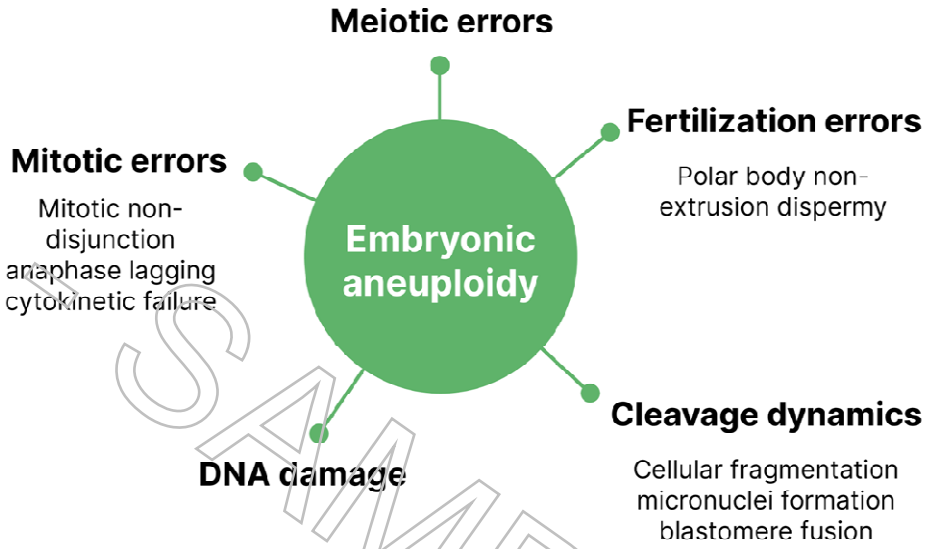
AIVF Genetics

A non-invasive AI screening tool for real time prediction of the embryo's genetic integrity



Introduction

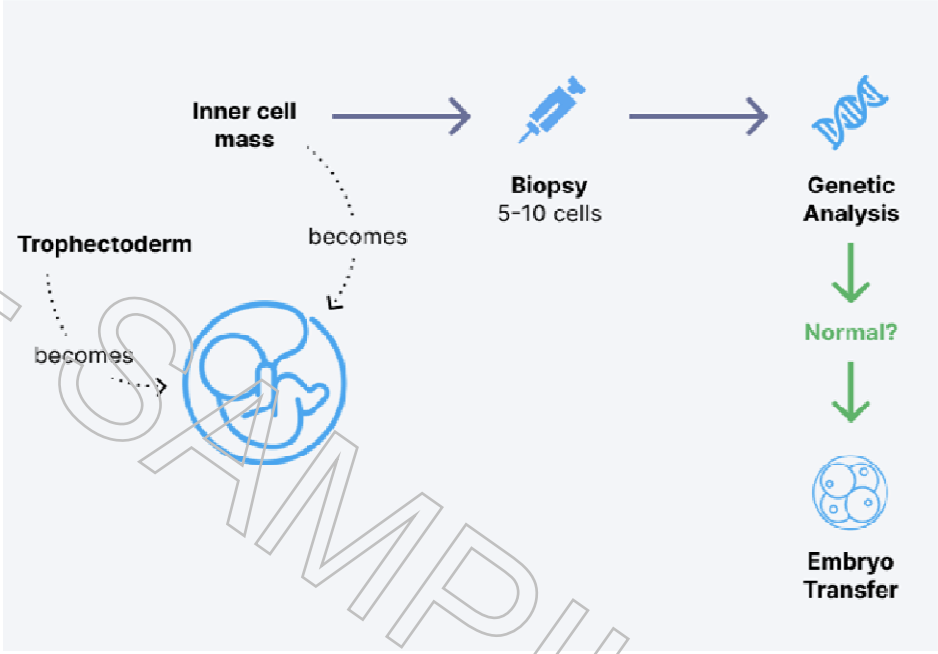
- Aneuploidy is the leading cause of recurrent implantation failure, miscarriage, and congenital abnormality
- Depending on age, 60-75% of preimplantation embryos that appear normal are aneuploid



šuiko O, Jatsenko T, Parameswaran Grace LK, Kurg A, Vermeesch JR, Lanner F, Altmäe S, Salumets A. A speculative outlook on embryonic aneuploidy: Can molecular pathways be involved? Dev Biol. 2019 Mar 1;447(1):3-13. doi: 10.1016/j.ydbio.2018.01.014. Epub 2018 Jan 31. PMID: 29391166.

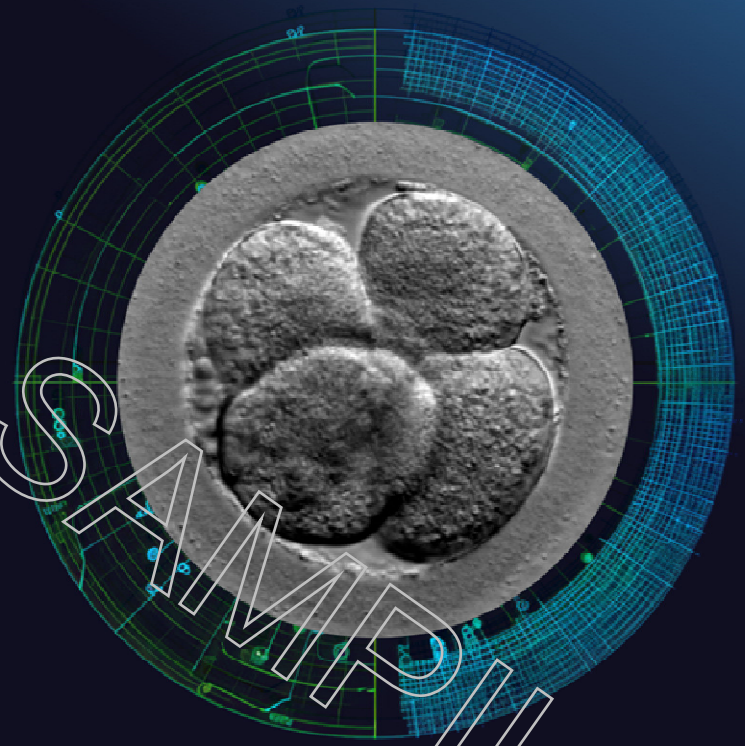
Introduction

- Euploid single embryo transfer is the central dogma of IVF success
- Current status quo for preimplantation evaluation: morphology + PGT-A
- Substantial limitations of PGT-A:
  - Expensive
  - Requires specialty embryology staff
  - Significant turnaround time to results
  - Embryos may not survive biopsy
  - Risk of cryogenic damage
  - Risk of false positive/false negative result
  - Unfit for patients with history of embryo damage
  - Unfit for patients who undergo fresh transfer



AIVF

There is demand for an alternative test that provides a quantitative embryo ploidy risk assessment.



The spectrum of ploidy testing

Understanding the difference between screening and diagnostic

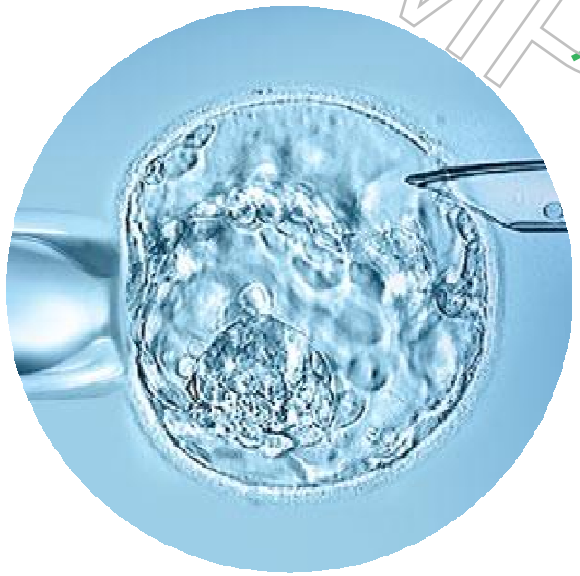
Screening

Risk scores are generated and modified based on biochemical analysis and population statistics

Diagnostic

Results are based entirely on genetic factors

What does a diagnostic test tell us?



Confirms the presence or absence of aneuploidy

Typically recommended for “high risk” embryos

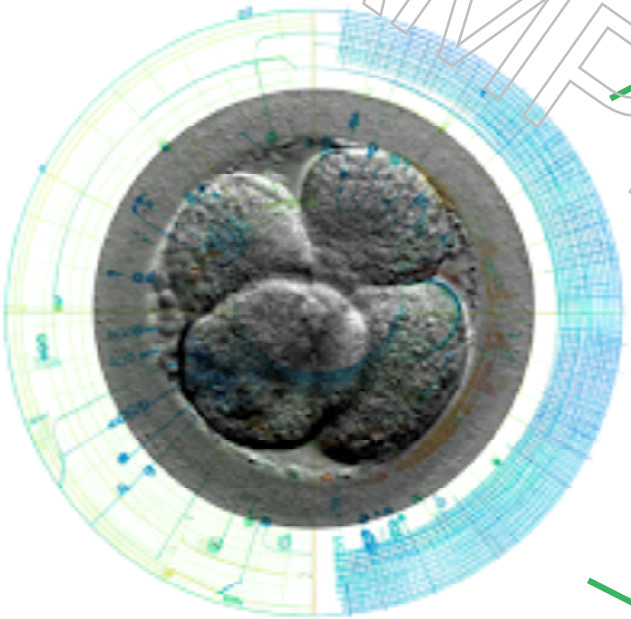
Diagnostic results should provide a yes/no answer to the existence of aneuploidy with as much certainty as possible.

Strength of the diagnostic test is determined by its accuracy

Cons: higher physical risk to the embryo + financial burden + time-consuming + high turn-around-time for results

AIVF™

What does a screening test tell us?



Intended to characterize genetic risk

Cutoff thresholds are used to identify embryo(s) that have a “high-risk” or a “low-risk” of aneuploidy

Can detect potentially abnormal embryos, while minimizing unclear or unknown results

Less invasive, therefore quicker to perform and less costly

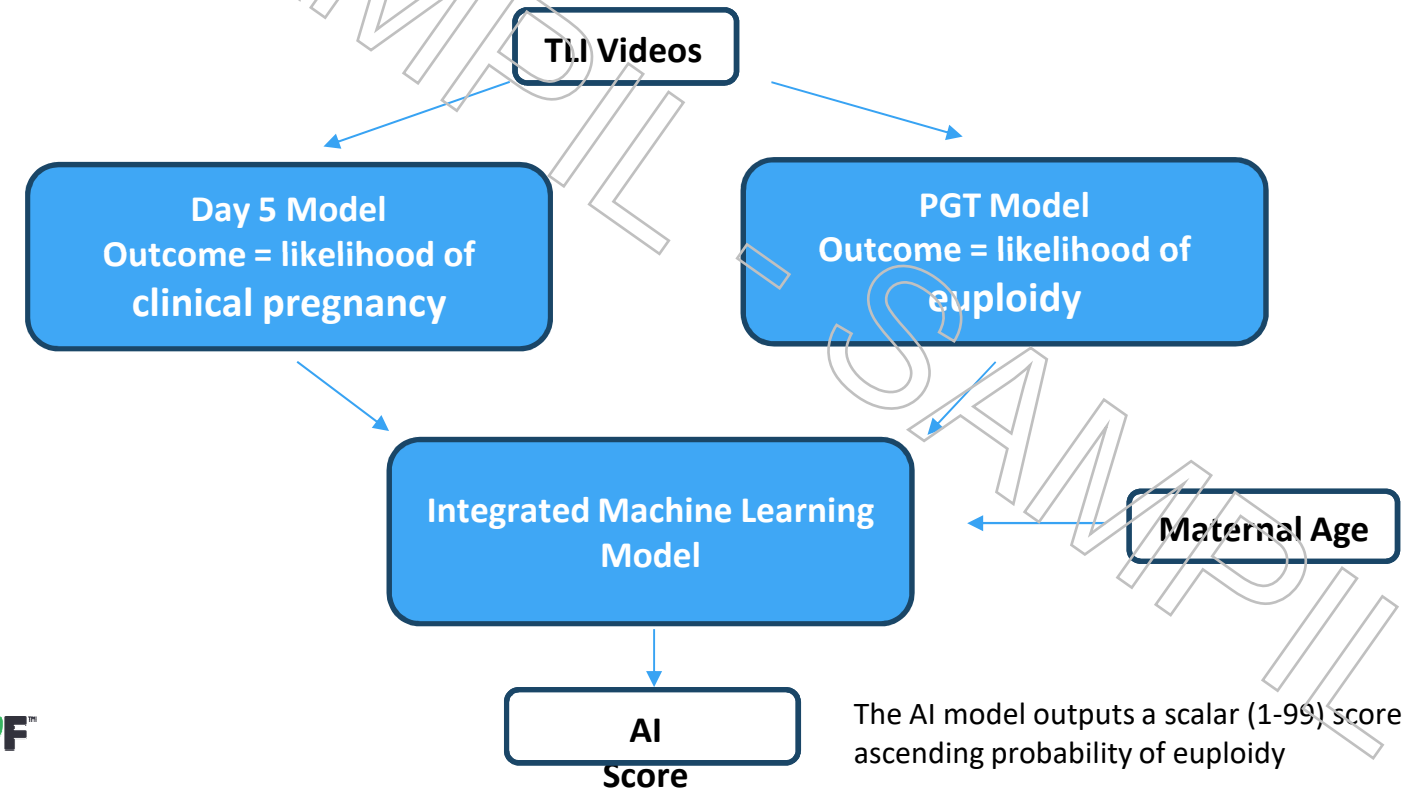
Not a diagnostic!

A “high likelihood of aneuploidy” result does not eliminate the possibility of that embryo being euploid.

AIVF™

Methods

Develop an AI-based genetic score that can provide real-time, reliable and biologically justified estimates of embryo ploidy (n=5,000 embryos)



The AI model outputs a scalar (1-99) score that associates with ascending probability of euploidy

Methods

We used a **confusion matrix** to characterize our screening test performance

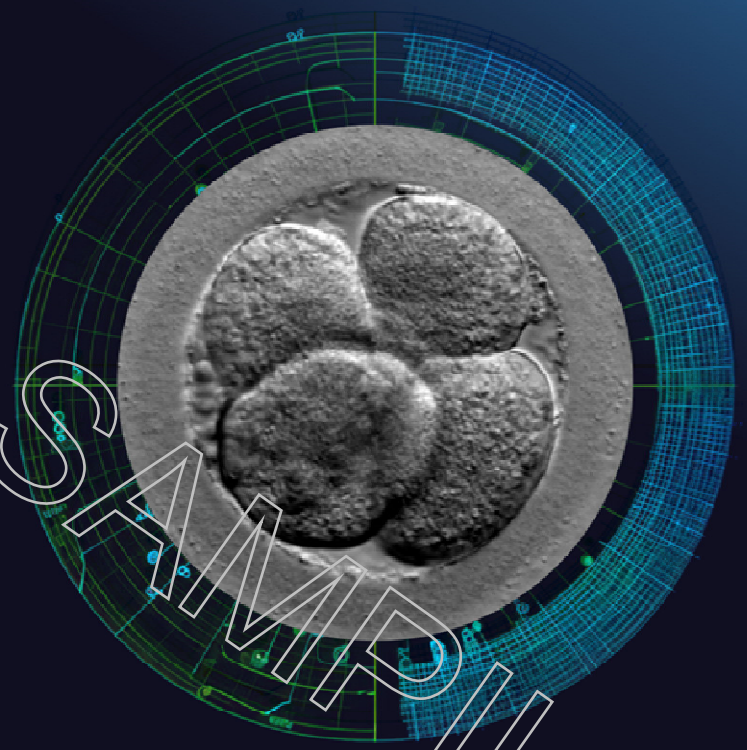
- The screening matrix generates positive/negative predictive values for every possible AI score threshold (AI scores range from 1-99)
  - The AI scalar increases with euploidy likelihood
  - **Positive result = euploid screening result)**
  - **Negative result = aneuploid screening result)**
  - False negative = true euploid embryos assigned an aneuploid label by the AI)

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- **The optimal AI score threshold for detecting/deselecting at-risk embryos with high probability of aneuploidy was determined to assess its most relevant clinical application.**

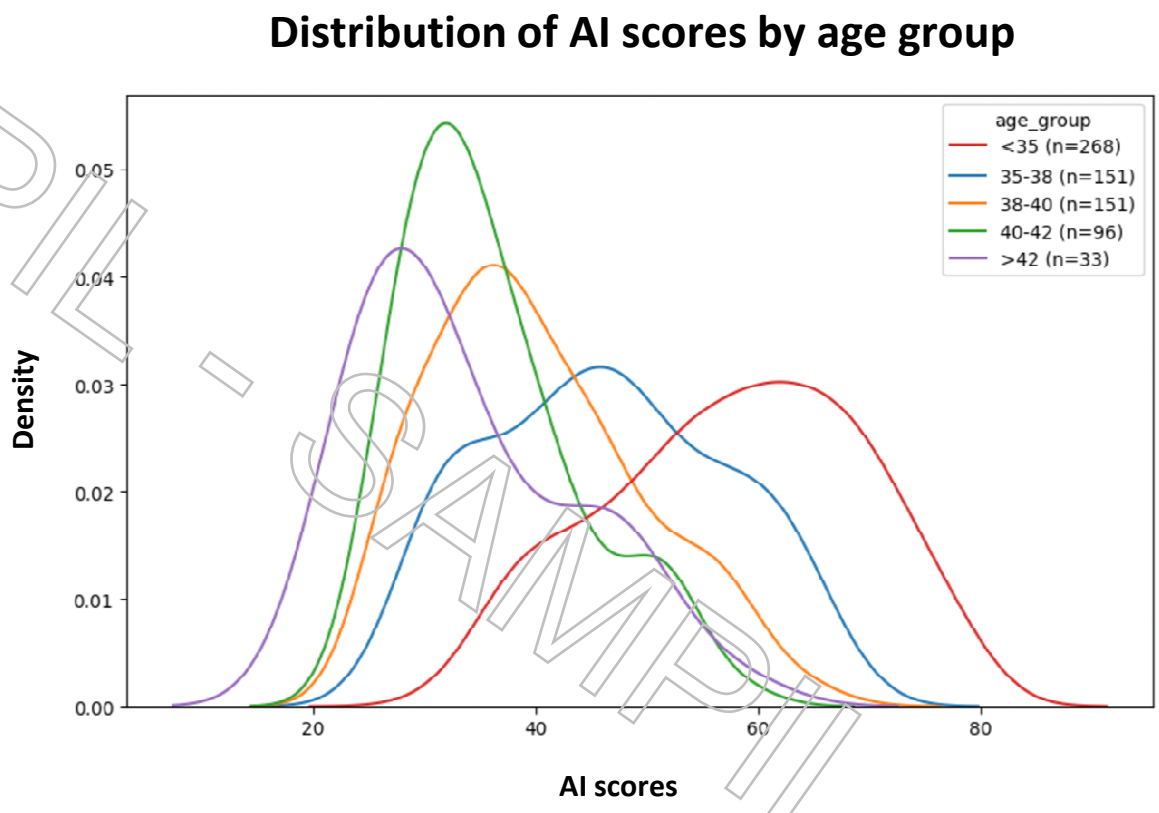
AIVF™

RESULTS



Results

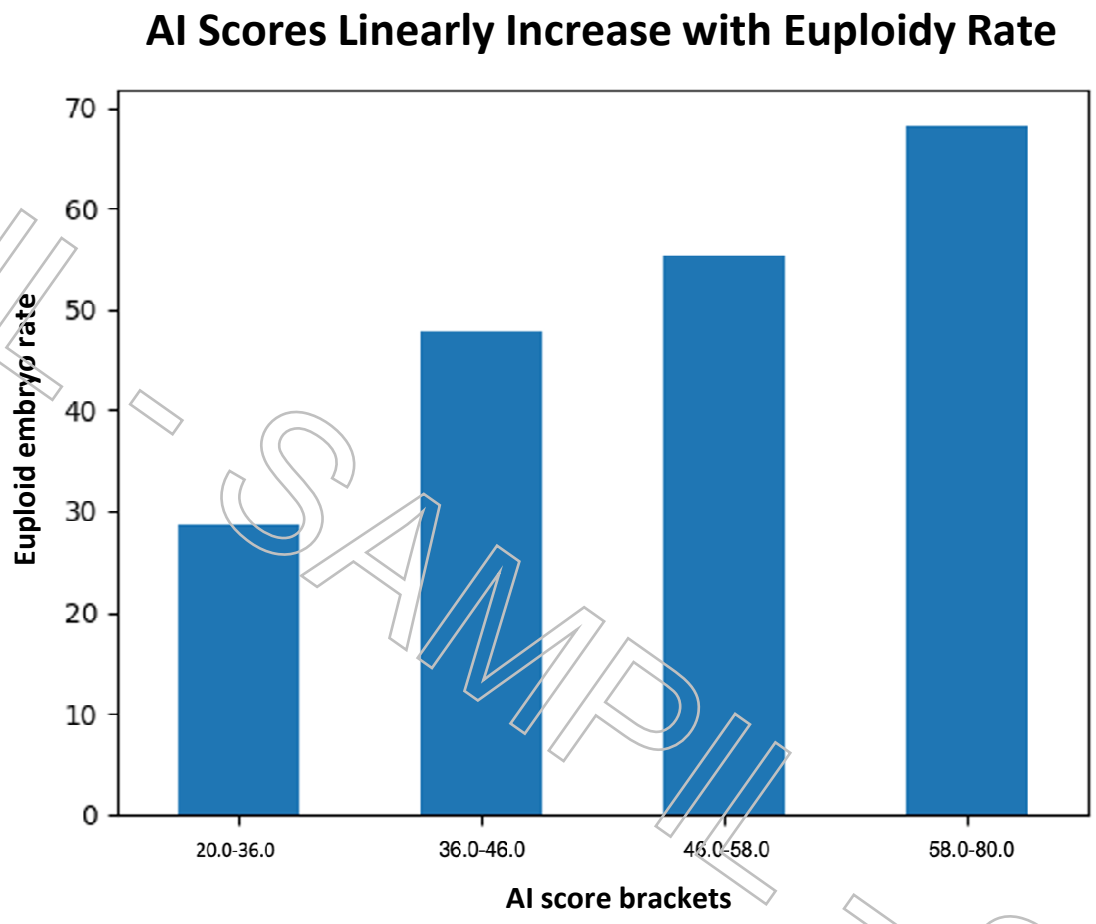
- Bell-curve distribution of AI scores shifts with increasing maternal age.
- Distribution of scores is statistically different between age groups.
  - (p value<0.001 was obtained using Mahn-Witney test)



AIVF

Results

- There is a significant linear relationship between increasing AI scores and ascending euploidy rate.
  - (p value<0.001 was obtained using Cochran–Armitage test)
- Odds ratio (OR) for the association between AI scores and euploidy probability = 2.79 [95% CI = 2.05-3.82].

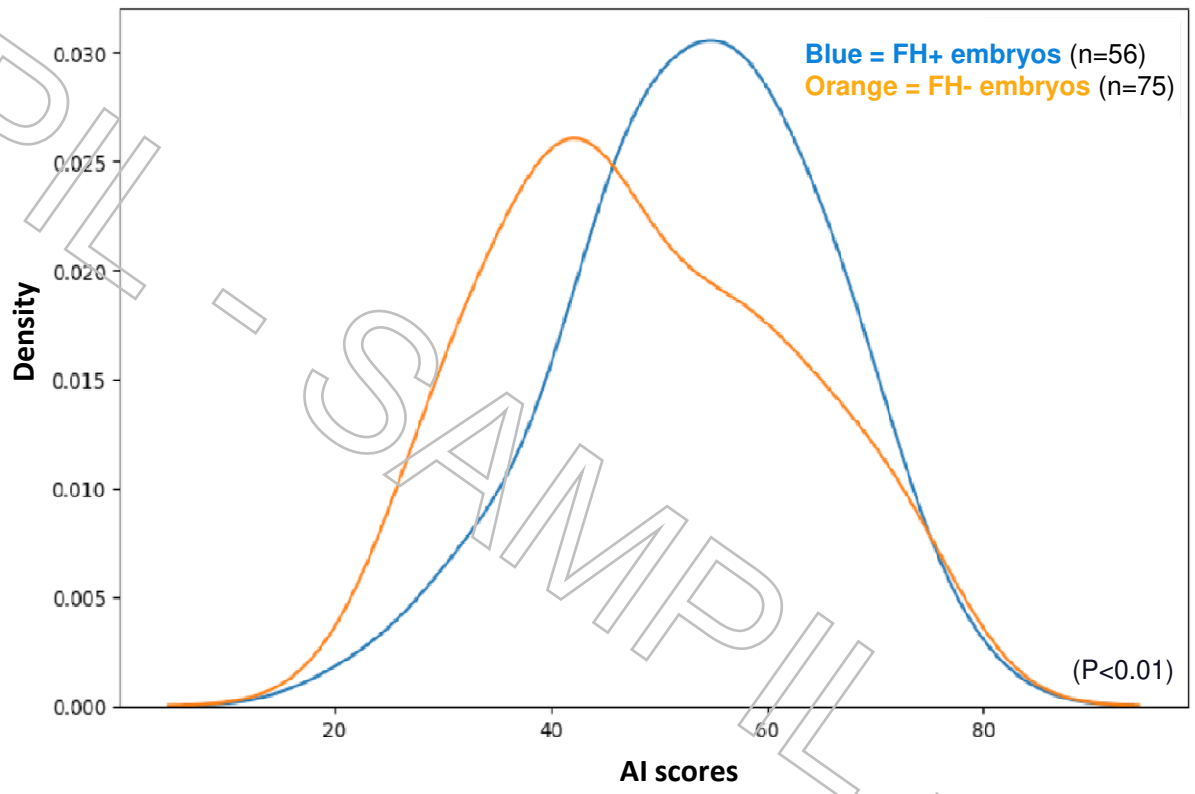


AIVF

Results

- Scores statistically discriminate between FH+/FH- embryo subgroups.
- FH+ embryos have higher scores than FH- embryos.
  - (p value<0.001 was obtained using Mahn-Witney test)

Distribution of AI scores by fetal heartbeat (FH) outcome



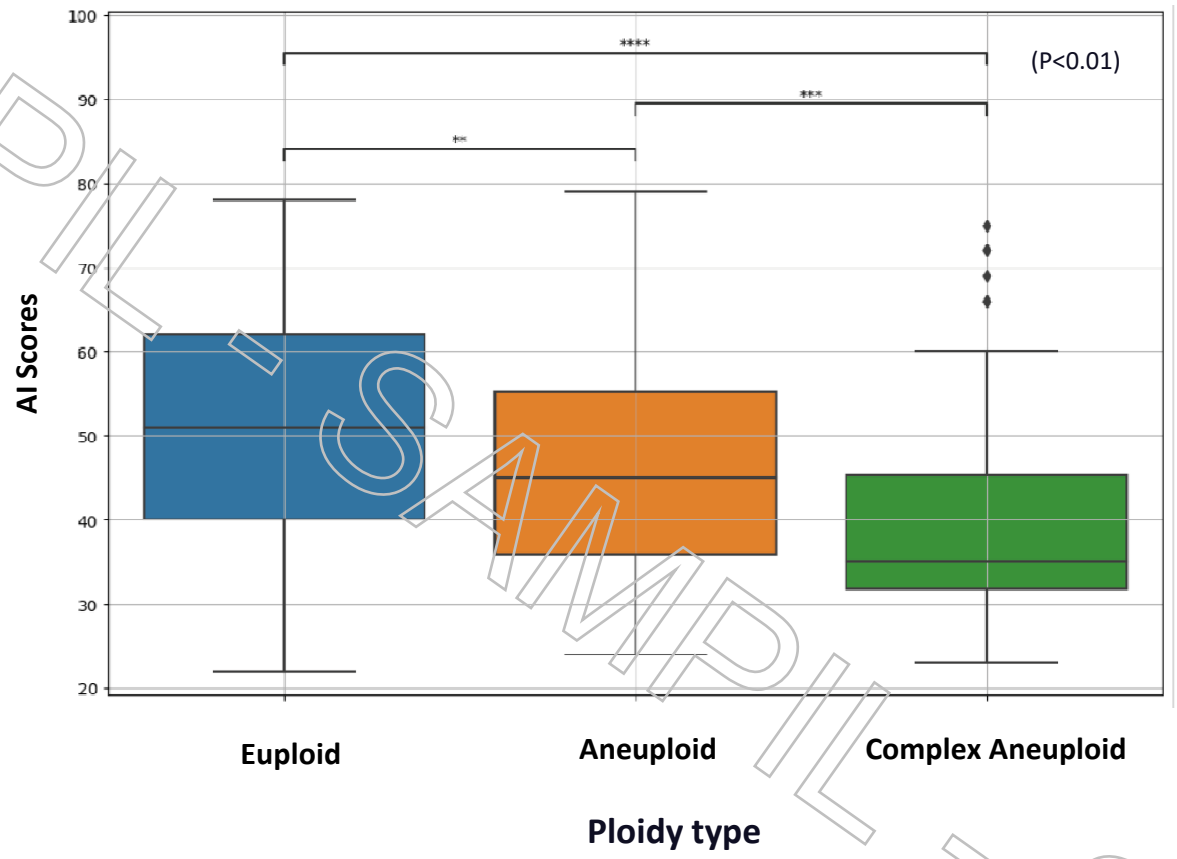
AIVF

Results

- Boxplot analysis shows the distribution of AI scores for euploid, aneuploid, and complex aneuploid embryos.
- The distribution of scores is statistically different between the different types of ploidy.
- The more aneuploid the embryo, the lower the score.
  - Pairs of categories - Student's T-test with Bonferroni correction
  - All 3 categories - ANOVA with Tukey's multiple comparisons post-test

AIVF

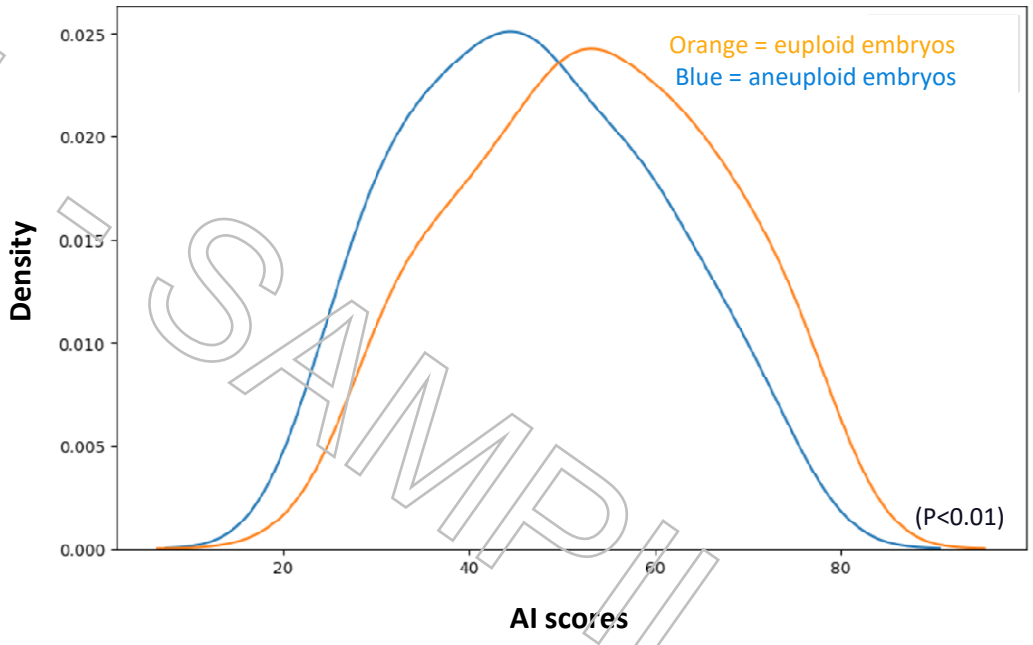
Distribution of AI scores for each type of ploidy



Results

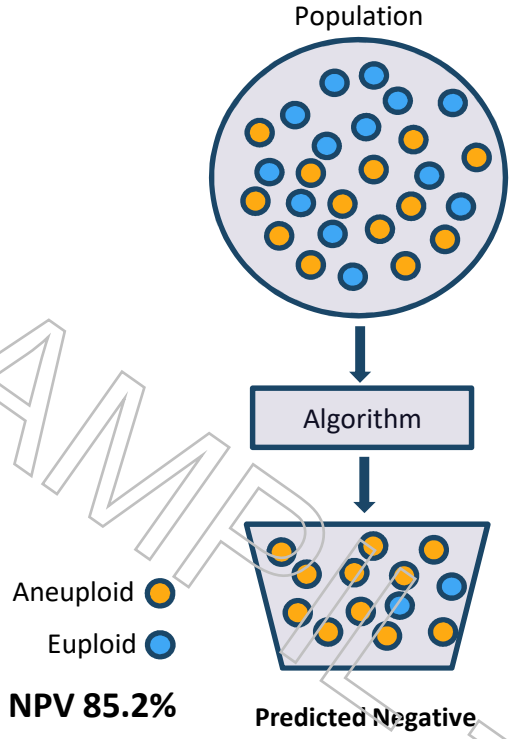
- Handling confounding factors: data inclusion: 66 patients with ≥ 2 embryos in their cohort; ≤ 1 aneuploid and ≤ 1 euploid embryo.
- AI scores discriminated between euploid/aneuploid subgroups per patient.
- **This analysis mitigates confounding variables due to selection bias; AI scores still showed robust performance.**
  - Selection bias = a common bias due to the fact that all embryos chosen for training have known outcomes and are not reflective of real-world cohort distribution data.

Distribution of AI scores by ploidy for patients with both ploidies in their cohort



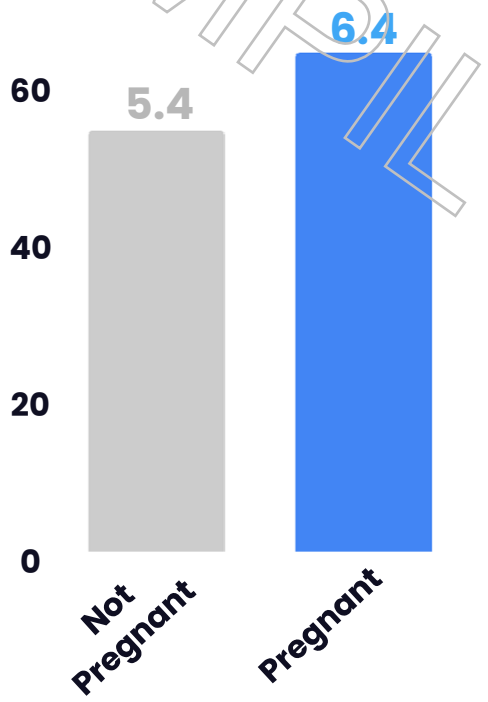
Results

- To demonstrate clinical utility, we determined optimal AI score of ≤ 28 for the confident **deselection** of embryos with high likelihoods of aneuploidy.
  - **Embryos that received an AI score ≤ 28 were correctly screened as aneuploid 85.2% of the time**
    - (this score had the highest negative predictive value [NPV] in the confusion matrix).
    - NPV= proportion of true aneuploid embryos that also had aneuploidy screening results.
  - The number of false-negative labeled embryos were markedly low (<9/669) below this AI score threshold.
    - False negative = number of true euploid embryos that had an aneuploidy screening result.

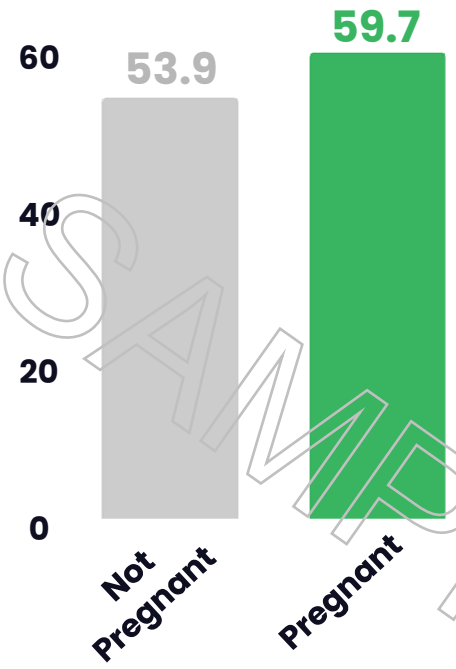


AIVF

AIVF Day 5 Scores per Clinical Pregnancy Outcome



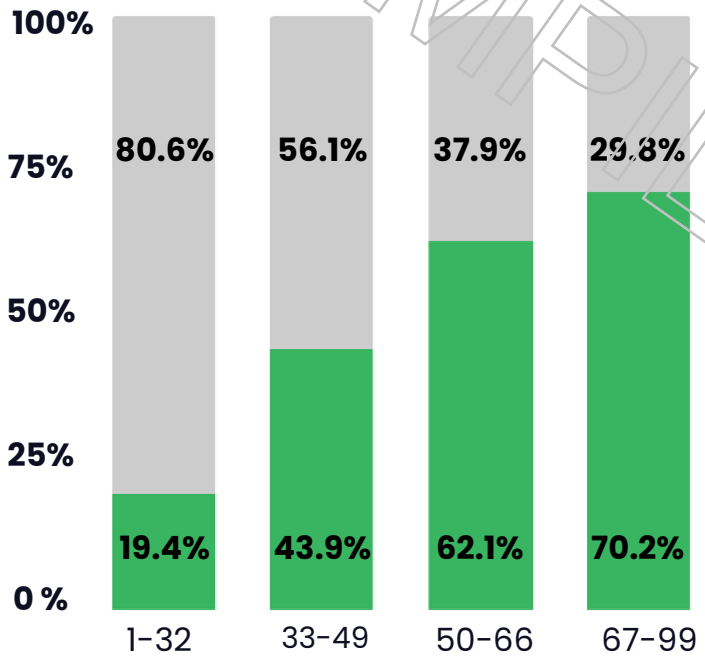
AIVF Genetic Scores per Clinical Pregnancy Outcome



AIVF™

EMA Genetics Provides Reliable Non-Invasive Screening

Correlation between AIVF Genetic Score and PGT Outcome



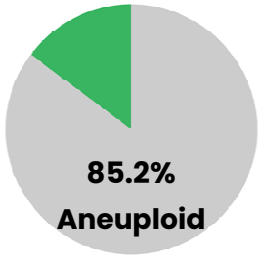
AIVF

Aneuploidy Rate

Euploidy Rate

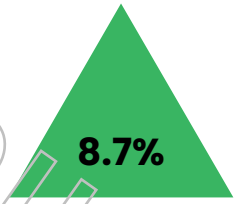
Accurately identify Aneuploid Embryos

Embryos with a Genetic Score ≤ 28



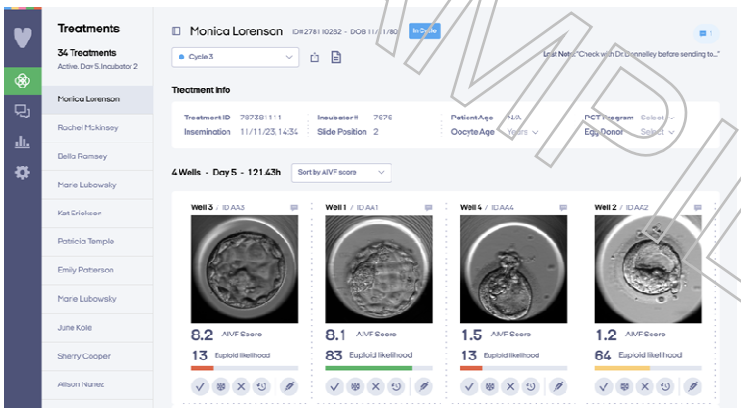
Confidently choose euploid embryos for transfer

EMA Score Difference in Aneuploid vs. Euploid embryos from the same patient



n= 1920 embryos (4 clinics)

Results



AIVF Genetics Score Brackets and Their Probabilities of Euploidy				
AIVF Genetics Score Bracket	High Likelihood of Aneuploid	Likely Aneuploid	Likely Euploid	High Likelihood of Euploid
	1-32	33-49	50-66	67-99
Probability of Euploidy (Expressed as an Averaged Percent)	28%	44%	58%	71%

- The highest level of model confidence was achieved at the tail ends of the scalar
- Embryo with a score above 66 were 2.5X more likely to be euploid than an embryo with a score below 33
- Importantly, exact probability estimates of euploidy per score bracket may vary per clinic, depending on the demographic and clinical practice.

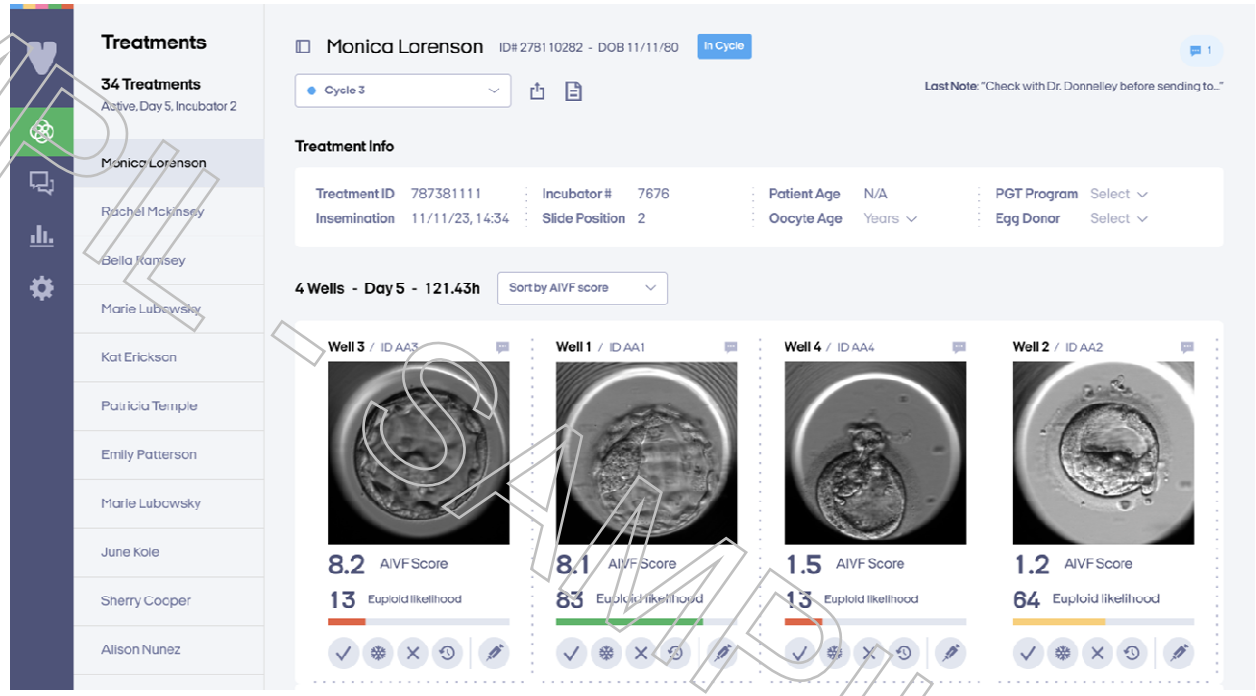
AIVF™

Results

For user deployment, four score brackets are defined for the user in the AI ploidy test interface:

- High likelihood of euploid: [67-99]
- Likely euploid: [50-66]
- Likely aneuploid: [33-49]
- High likelihood of aneuploid: [1-32]

AI ploidy screening test interface



AIVF

Clinical Benefits of a ploidy screening test

Shortened Time-to-Pregnancy:

- Facilitates rapid and effective screening of embryo genetic quality on Day-5 without the need for results turn-around-time.

Robust Alternative for Unfit Candidates:

- Serves as a valuable alternative for patients unsuitable for invasive preimplantation genetic testing.
- Provides crucial embryo quality information to patients who will not undergo invasive procedures or those undergoing fresh embryo transfer.

Optimized Operational Efficiency and Reduced Costs:

- Reduces the workload and operational expenditure of IVF labs by minimizing the necessity for highly skilled embryologists to perform laborious diagnostic procedures.

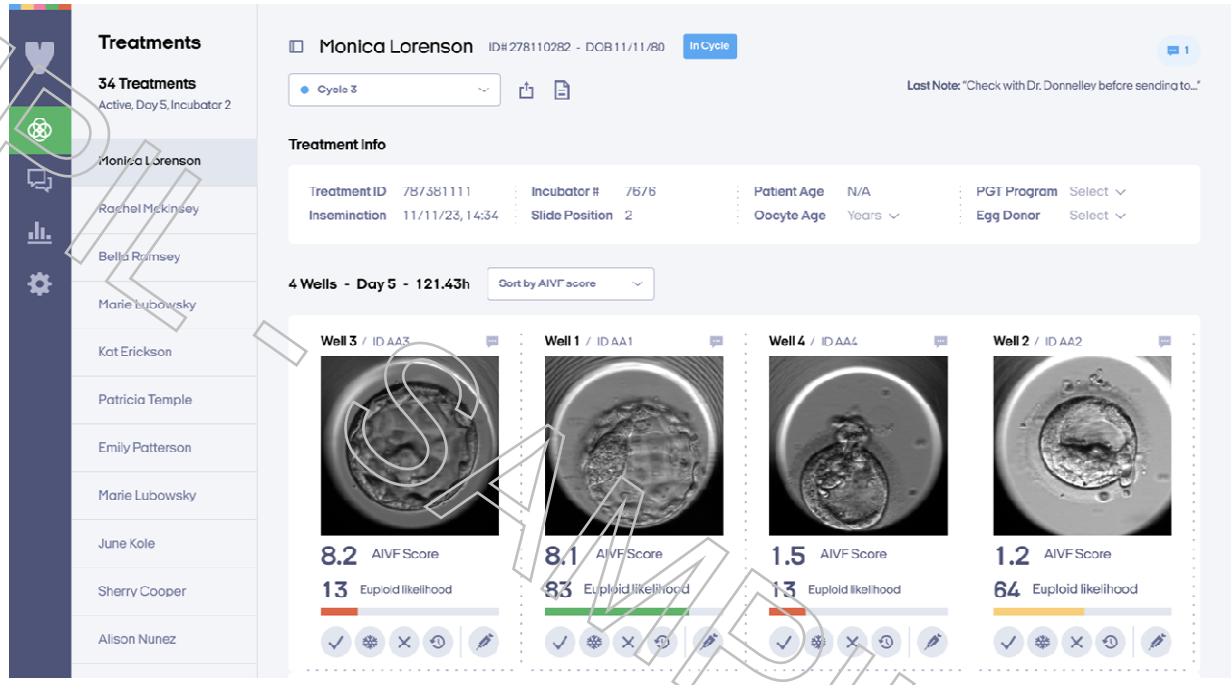
Enhanced Patient Prognosis Counseling:

- Drastically optimizes patient counseling by offering prognostic forecasts of embryo genetic quality swiftly.

AIVF™

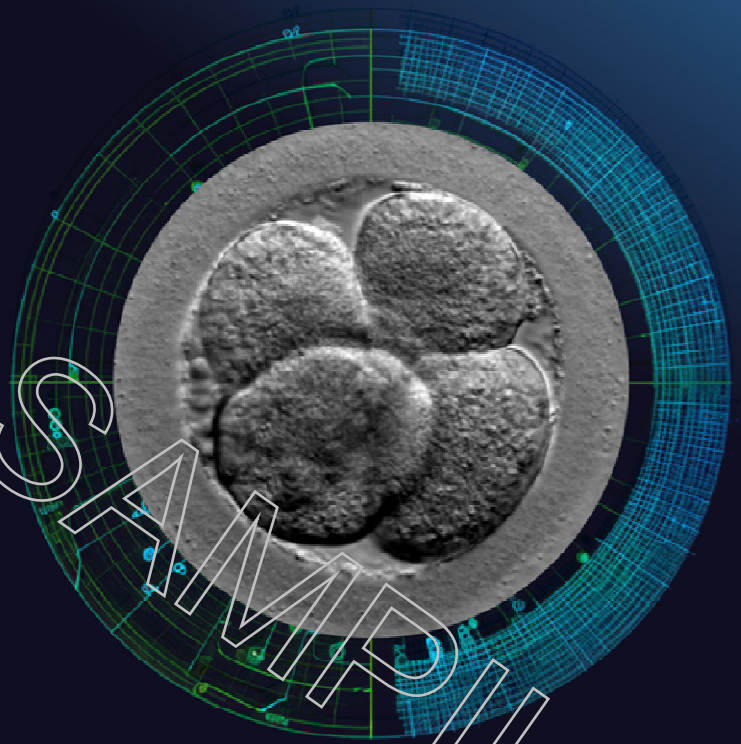
Limitations

- The screening does not provide testing information on the chromosome level.
- The influence of mosaicism was not assessed.
- This is not a diagnostic test.



AIVF™

Our Team



Our amazing AI & clinical Team

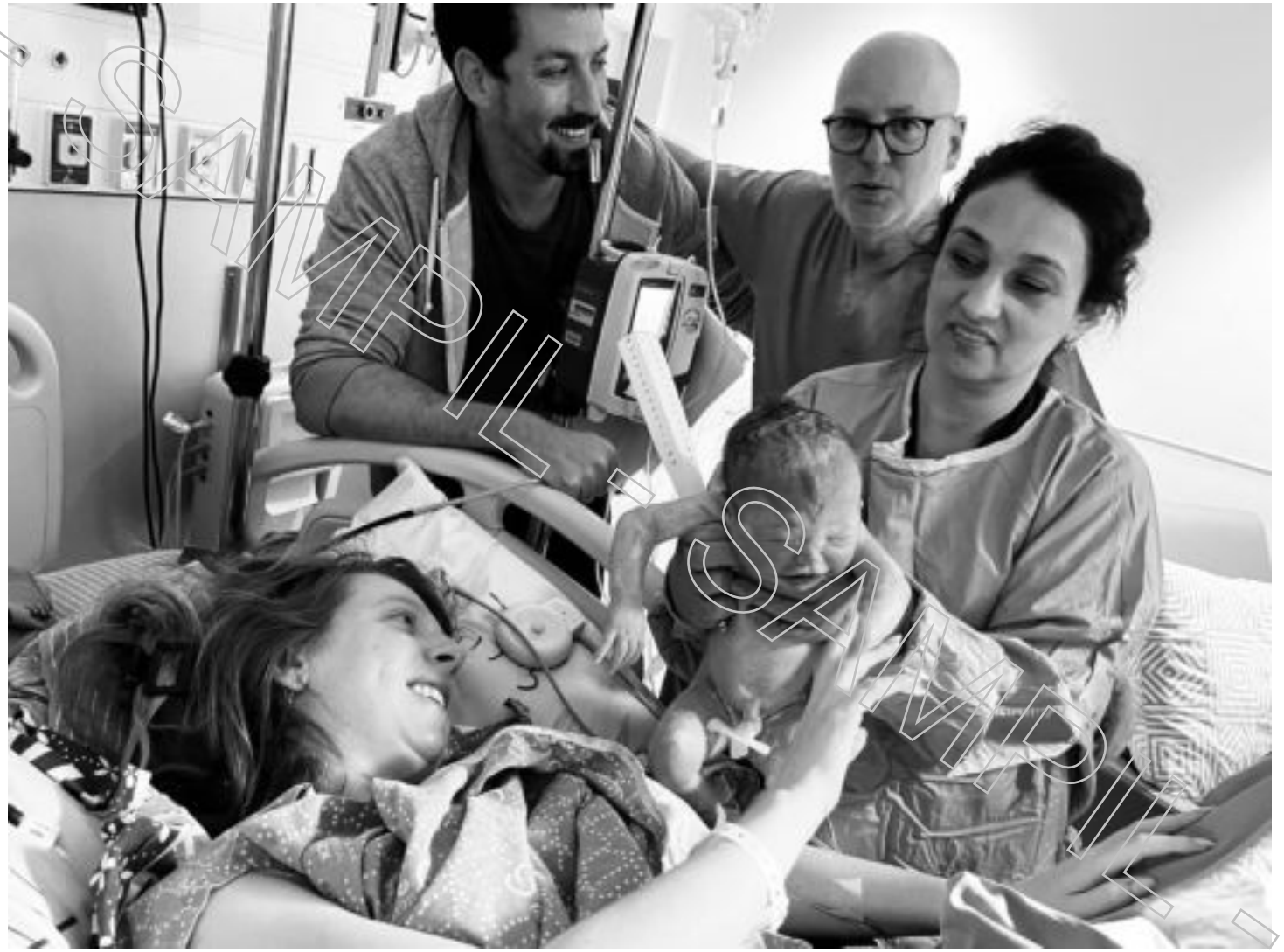
- Yishai Tuaber
- Itamar Tsayag
- Amichay Feldman
- Ron Amir
- Yuval Amar
- Yonatan Paserman

- Maya Shapiro
- Nicole Lustgarten
- Tamar Schwartz
- Michal Shelef
- Dan Coster
- Eyal Ohayon

AIVF™

Future direction

- Beyond embryo ranking: dependable pregnancy probability estimations for more transparent, dependable AI embryo evaluation
- Future studies should be directed toward evaluating clinical agreement between model predictions and observed outcomes



AIVF™

THANK YOU

AIVF™

NEJM AI

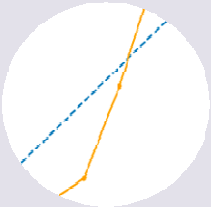
Published September 26, 2024
NEJM AI 2024; 1 (10)
DOI: 10.1056/AIp2400583

PERSPECTIVE

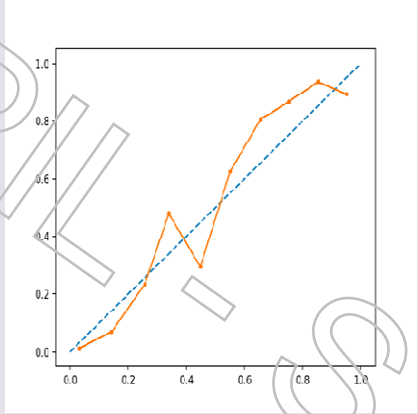
Settling the Score on Algorithmic Discrimination in Health Care

Marzyeh Ghassemi, Ph.D.,[1,2] Maia Hightower, M.D., M.P.H, M.B.A.,[3] and Elaine O. Nsoesie, Ph.D.[4]

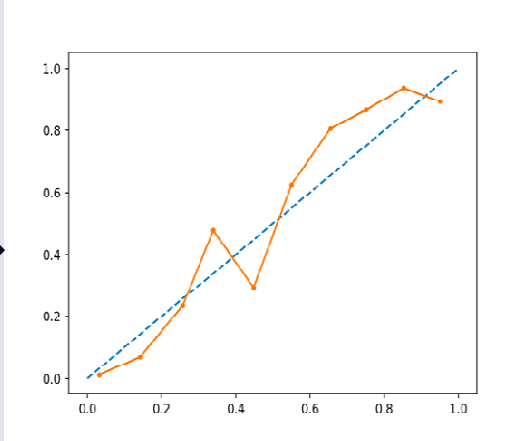
- Creating and funding quality assurance laboratories that have diverse, local, deidentified data sets could be a path toward both independently validating models for absence of bias and enabling researchers to accelerate the development of improved models.
- Importantly, a model that performs optimally in one health setting — that is, that balances best overall patient performance with the lowest fairness gap for any group — is not guaranteed to perform optimally in other settings.
- Thus, **local validation is necessary before models can be used on a population.**



Computation of top clinical variables most important for model prediction

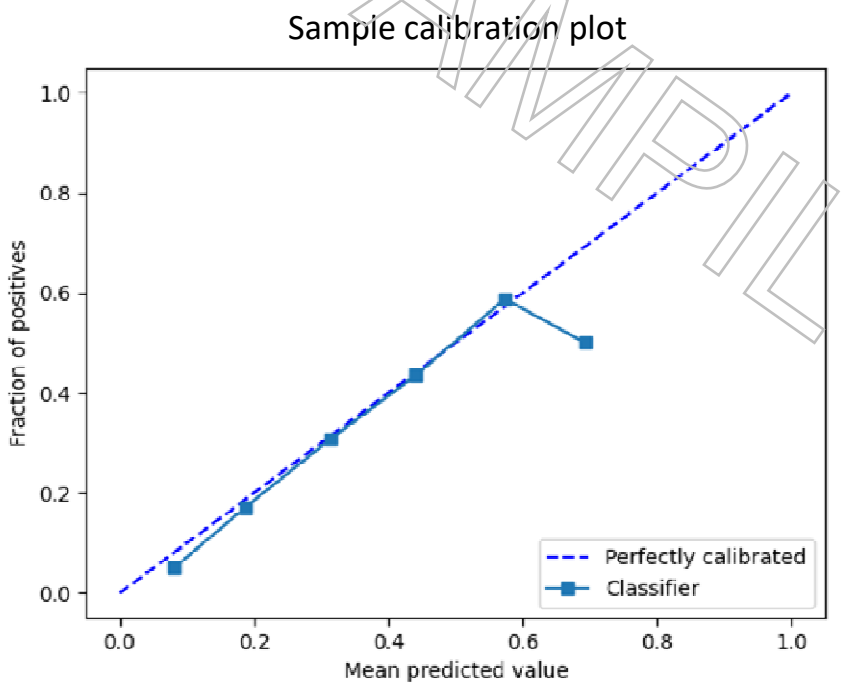


Further data processing into final ML prediction model



ML calibration per-clinic-dataset

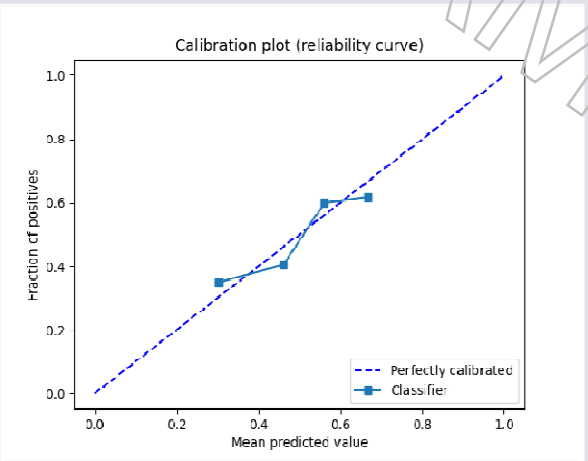
Quick guide on model calibration



The better the calibration, the closer the plot curve is to this straight line.

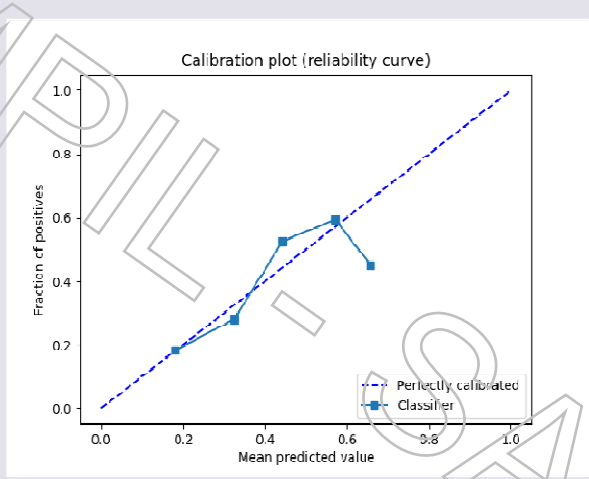
- Shows potential mismatch between the pregnancy probabilities predicted by the model, and real probabilities observed in the clinic data
- This model overestimates when predicting low probabilities of pregnancy and underestimates when predicting high ones.
- For samples for which the model predicted the possibility of being positive to be around 30%, only 10% of them were indeed positive.
- Conversely, almost all samples with predictions of 90% were positive.
- This model is not calibrated!

Clinic 1
Spain (n=660)



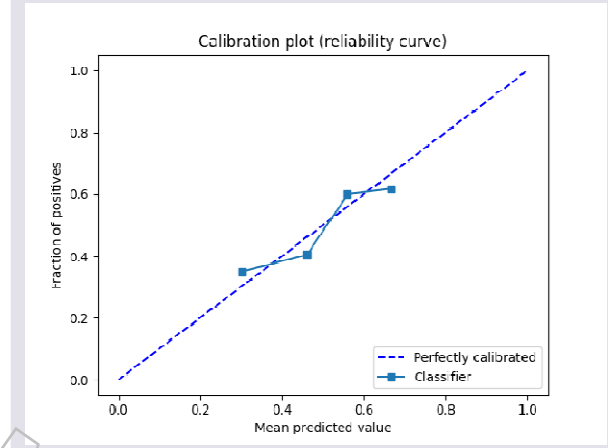
Brier loss score: 0.24

Clinic 2
Israel (n=345)



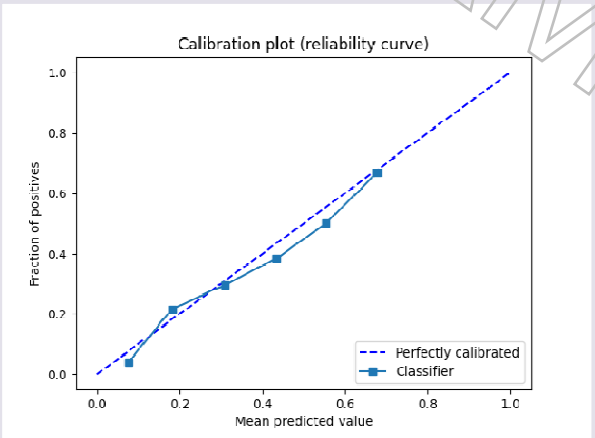
Brier loss score: 0.17

Clinic 3
USA (n=300)



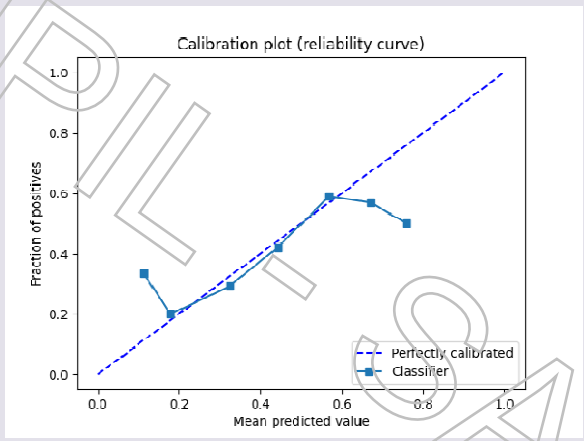
Brier loss score: 0.24

Clinic 1
Spain (n=6600)



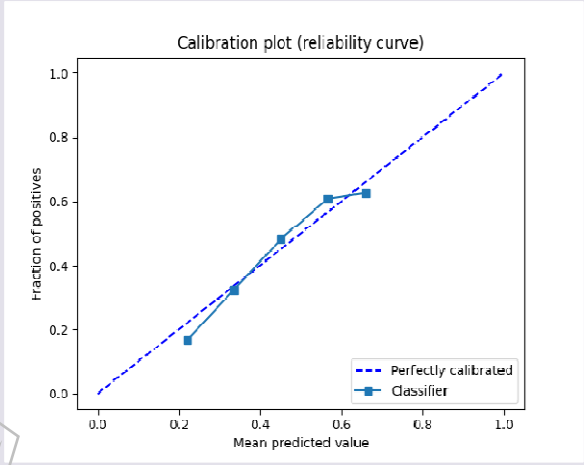
Brier loss score: 0.24

Clinic 2
Israel (n=1700)



Brier loss score: 0.17

Clinic 3
USA (n=1490)



Brier loss score: 0.24

Conclusions

- **AI accurately accounts for clinical variables likely to influence its predictions**
  - Indicated by the modest contribution of SHAP analysis to the overall AI performance
- **Model calibration on a per-clinic basis is recommended for enhanced ML performance and score interpretation**
- **Calibration is an important, overlooked aspect of AI model training. Our calibration framework enables a more reliable, personalized approach to embryo evaluation**

Individual patients can be given an accurate estimation of their pregnancy odds using AI-based embryo evaluation and patient metadata.