Hype or reality? Is artificial intelligence (AI) solving the real problem in the IVF laboratory?

Authors: Juliana Simões[1], Inês Couceiro[1], Maya Tsarfati[2], Daniella Gilboa[2], Daniel Seidman[2], Vladimiro Silva[1]

[1]Ferticentro SA, Coimbra; [2]AiVF, Israel

Introduction

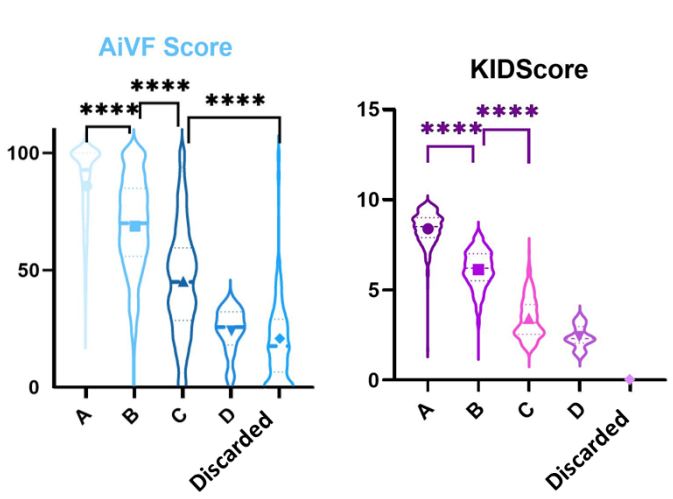
The impressive performance of AI tools that distinguish between top/good/poor embryos for implantation prediction has been shown in numerous published reports. We hypothesize that this does not reflect the AI's "real-world" efficacy; characterizing AI performance on a dataset using top/poor embryos that would otherwise be manually selected/deselected skews its presentation of efficacy and benefit. To reflect true clinical utility, the performance of AI tools should be assessed using a cohort of homogeneous B-grade embryos – i.e., the subset of embryos most relevant for AI decision-support. In this context, we compare the discriminative performance of EMA (AiVF, Israel), an AI platform for embryo evaluation, and another commercially available embryo scoring tool.

Methods

We examined the distribution of scores, through their averages and quartiles, for each category: top, good, fair (A, B, C-grade embryos, respectively). We then focused on the subset of embryos that were not top/poor quality (i.e., B-grade embryos). Data was analyzed by Student's t-test (Python Script Mode). Differences were considered significant at $P < 0.05$.

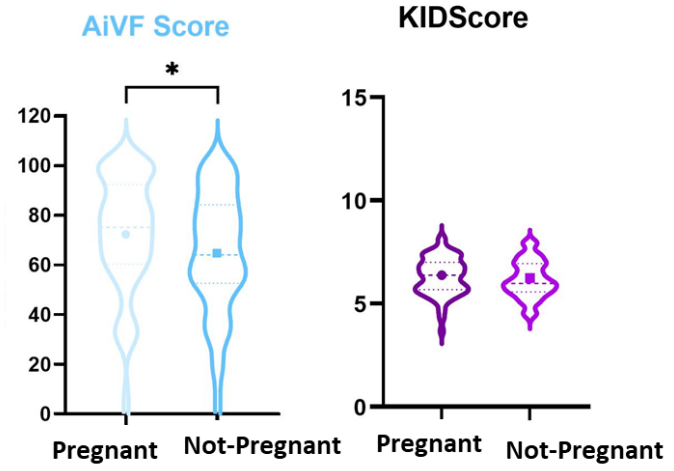
Embryos (fresh or vitrified-thawed) with known implantation outcomes	647
Number of patients	500
Number of treatment cycles	541
Grade-A embryos	667
Grade-B embryos	582
Grade-C embryos	753
Grade-D embryos	10

Results



	A	B	t_{df}	p
Mean **AiVF Score** (1 - 100)	86 (Std 17)	68 (Std 21)	15.49	<0.01
Mean **KidScore** (1 - 9,9)	8.4 (Std 0.92)	6.13 (Std 1.06)	39.99	<0.01

Both models showed statistical differences between mean scores for grade-A/grade-B embryos ($p<0.01$), demonstrating robust ability to distinguish between morphology classes.



	Pregnant	Not-Pregnant	t_{df}	p
Mean **AiVF Score** (1 - 100)	72 (Std 24)	64 (Std 24)	2,12	0,03
Mean **KidScore** (1 - 9,9)	6.34 (Std 0.92)	6.20 (Std 0,86)	0,97	0.33

When examining the distribution of scores for B-grade embryos only, EMA showed statistically higher scores for the B-grade implanted group, compared to the B-grade nonimplanted group ($p=0.03$). This difference was not observed for the commercially available tool ($p=0.33$).

Conclusions

When focusing on the subset of embryos most relevant for AI analysis, EMA more accurately differentiated between implanted/nonimplanted subgroups and reflected the KID. EMA has the potential to offer clinical benefit to enable objective, accurate embryo evaluation decision-support.

Acknowledgements

The authors would like to acknowledge the valuable contributions of their colleagues in both Ferticentro SA and AIVF. In particular all the embryologists that devoted their time to help with the classification and by doing the manual annotations required for the morphokinetic tool.

Conflict of interests

EMA is a product of AIVF, Israel. Ferticentro SA has no connection to this product and is an independent element of this study.

Fertility2023
10-13 January 2023
ICC Belfast

AIVF