ESHRE 2020 Virtual (5-8 July 2020)

Questions for the speakers

Session 31: Predictive algorithms in Clinical Embryology

Artificial Intelligence (AI) system combining both images and non-image inputs can improve the accuracy of human embryo viability prediction - Isao Miyatsuka (Japan)

Q: Why did you evaluate Ca-125, a cancer marker?

A: We mentioned the sample of patient information like some hormonal value and CA-125, but in this analysis we do not use CA-125.

Q: How did you assess the influence of menstrual cycle?

A: We use this features as the number of day in particular menstrual cycle.

Q: What's the reason for choosing positive pregnancy as the primary outcome in most AI models, and not live birth?

A: Of course, We know the importance of live birth, but this time we do not have enough amount of clinical case of live birth. So we set positive pregnancy as the primary outcome.

Camera-agnostic self-annotating Artificial Intelligence (AI) system for blastocyst evaluation - Matthew VerMilyea (U.S.A.)

Q: Have you adjusted the prediction of embryo viability to female age?

A: No, the prediction of viability is not manually adjusted at any point. The proportionate contribution of factors that affect the viability prediction score are controlled for implicitly by incorporating a representative demographic at the time of AI training. This is tested by performing an age subgroup analysis on a test dataset. It was found that viability prediction in a younger (< 35) demographic is much more likely to contain False Positives (a high viability prediction but unsuccessful clinical pregnancy) linked to patient factors, for patients undergoing IVF.

Q: What is the general AUC of the prediction model and given blast dynamics: which time-point was used for standard images

A: For the Cleaned dataset, the AI Model predicts with an AUC value of 84% versus the equivalent embryologist value of 67% on a dataset where a portion of patient factors have been controlled for (removed).

On the full (uncleaned) clinical dataset with patient factors preserved (938) images, the AUC value is 61% compared to an embryologist value of 55%. Note that the AUC is not the best measure of performance for datasets where the total numbers of the viable and non-viable examples are unequal.

The time point used was Day 5 after IVF, which is constrained to within a 24 hr time window. The closest time point available prior to implantation was used.

Q: How can we account for different resolutions and illumination settings on an AI model from different TL systems?

A: The embryologist should not be required to account for different resolutions and illumination settings, as the AI was trained in such a way as to control for a range of real-life resolution and illumination scenarios from a range of different camera setups. The training also involves extraction of certain key morphological features, and the cropped/centered blastocyst itself, which are robust to image artifacts (such as the TL system 'well' shape). It was found that the change in resolution and illumination/brightness of the image caused little different in the AI prediction score, so long as the image was not distorted to the point of loss of extractable information.

Q: Are the standard microscope images for all embryos taken at a fixed time post insemination?

A: The standard microscope images used for training and testing of the AI were sourced from real clinical data, which covers a range of time points within a fixed 24 hr window, covering Day 5 after IVF. No images outside this time were considered admissible in the study and intended use of the AI algorithm. However, a range of times within Day 5 after IVF were considered. Outside of this study, our AI has shown similar accuracy for images through Day 6, however the intended use and most of the testing has been for Day 5.

Q: Regarding cleaning of data, could that introduce inherent biases from the cleaning criteria? How did you mitigate for this?

A: The cleaning process, while mentioned briefly (to be more thoroughly described in an upcoming poster presentation at ASRM 2020), was introduced as a measure to reduce bias, occurring from inherent errors in the data when viable embryos are labeled as non-viable when non-pregnancy is due to patient factors (e.g. endometriosis) and not the embryo quality.

The patent-pending UDC method has been tested extensively in a range of settings with known ground-truth outcomes, and has been shown to be a robust technique for identifying mislabeled data (where 'mislabeled' here indicates that measurement of clinical pregnancy after six-weeks is imperfectly correlated with viability in certain specific counter-examples). For example, it has been applied to clean the training data to create our AI, and shown an improvement in accuracy and generalizability on both cleaned and uncleaned test datasets.

The removal of these counter examples using the UDC was handled in a data-agnostic way so as not to introduce positive bias toward model performance, beyond the handling of patient factors and other causes of data mislabeling. The UDC method involves a specific AI method which is separate to the AI embryo model.

Q: >600 images out of 900+ were excluded from analysis. What is the value of your AI system in a real clinical setting if only 1/3 of embryos can be assessed?

A: This is incorrect. All of the data are assessed in this study, and all data are capable of being assessed in the system. While a full description of the UDC cleaning method did not lie within the scope of this particular presentation, we present two results of the AI model performance: one case where the total clinical dataset including patient factors was preserved (Blind Test Set 1 – 938 images), and additionally, report on a subset of the data where likely-patient factors had been identified using an independent method (Blind Test Set 2 – 696 Images of the 938 Images). Both reports represent valid results, since the presence of patient factors simply increases a portion of False Positives where a potentially ideal or viable embryo (predicted by the AI, or by the embryologist) does not result in a clinical pregnancy for reasons other than embryo viability.

Q: Are the data (hpi) of blastocysts images taken (when no time-lapse used) and taken into consideration in the model?

A: Data pertaining to the blastocyst such as hpi are only implicitly taking into consideration together with other overlapping factors in the morphology, and the proportion or weight of each part of the morphology extracted from the image that contributes to viability, is optimized during the AI training process. Therefore, there are no hand-optimized or manual changes to the consideration of the components of the image even though it is implicitly contributing to the training process.

Q: Few patients have several similar blastocysts in their cohort to be selected, do you think that AI evaluation will be needed for many IVF cycles? ??

A: AI evaluation is difficult to test within cohorts because the endpoint for each embryo in the cohort is not known (not all embryos are implanted). Therefore, the effect of the AI on cohort ranking can only be assessed either in a longitudinal study measuring average time to pregnancy, or it can estimated using an innovative method to be presented in an upcoming poster at ASRM 2020. Ideally, AI evaluation should be tested both in-clinic (across many cycles) and in studies that have a fixed measurable end point for analysis.

Q: Blastocysts are highly 3D and dynamic. ICM could be visible in non-central plane. Which single image from TLM you consider representative?

A: It is indeed the case, that there is a choice of focal plane, and the images considered cover a range of different focal planes across different embryologists. However, the most representative focal plane is the plane which intersects the ICM half way, so that the ICM is most in focus. In cases where the focal plane deviates from this point, we discovered there was not a significant change in the results of the study, since the blastocyst image still contains all the information (albeit, with reduced focus on certain features).

Q: How many embryos and patients were in the dataset and what was the distribution of the ground truth outcomes (fetal heartbeat?)?

A: The breakdown of the dataset is as follows: 3,689 Images of Day 5 blastocysts were used, from 3,112 patients, with 2,530 Images used for training and validation, a blind set of 938 Images was used (with 696 Images comprising the Cleaned test set), and 221 Embryo-scope Images used for testing the

AI model performance on time-lapse images. These include 1,979 embryos that resulting in a fetal heartbeat, and 1,710 embryos that did not result in a fetal heartbeat.

If the score that calculated by AI is low and the patient asks for clarification, how would you explain it?

A: In the case of the Life Whisperer AI, if the score calculated by the AI is low, this means that the AI is confident that the embryo morphologically is very similar to embryos that typically do not result in a clinical pregnancy. The AI is trained on a historical dataset, and this is an accurate way of describing the meaning of the score. Additional clinical information about the patient as known by the clinician and embryologist together should be taking into account when deciding which embryo to implant, as the Life Whisperer AI is intended for decision support.

Do you share AI data and the progress of AI studies with patients? Do you present/consider AI as an add-on? ??

A: Yes. Life Whisperer regularly publishes studies for both clinics and patients, to provide confidence and the clinical benefit of AI for improving IVF outcomes. Life Whisperer also considers AI as an add-on for clinics and patients. The Life Whisperer team believe a pay-per-use fee is the best approach, so that clinics do not have to pay any software licenses to access the AI for their patients (access to the software is free), and patients pay if they elect to use the AI to support their IVF treatment. This help ensure that AI can be accessed by any clinics and patient that wants to use it, with minimal barriers.

Q: ALL SPEAKERS: most algorithms predict outcomes on a per embryo basis, not per patient. Shouldn't clinical relevance be tested on patients' embryo cohorts?

A: It is true that there is value in a study across patient cohorts, however, there is no ideal method for testing it in this regard – because it is not possible to transfer all embryos to assess their outcomes for the clinical assessment.

A randomized study can be conducted, however there is no clear measurable end point for each embryo (as not all embryos are implanted), and therefore the useful measurement that can be conducted in this case are around average improvement of time to pregnancy over time.

As an alternative, we have designed an innovative cohort study that will be presented by Ovation Fertility and Life Whisperer at ASRM 2020 in an upcoming poster presentation.

There is nevertheless value in a per-embryo study as well, because the AI compares embryos on an individual basis. In this case, there is a measurable endpoint, such as total accuracy of prediction, balanced (weighted) accuracy across viable and non-viable embryos, sensitivity, specificity, positive predictive power, negative predictive power, Area-Under-Curve (AUC) and Precision-Recall (PR) curves, for example.

Q: ALL speakers presented results of efficacy of their AI with different detail. What do you think is minimum list of result types required to demonstrate efficacy?

A: The efficacy of an AI model will involve a measurement of the AI performance (as defined based on a measurable endpoint) on a double blind (or external) dataset, or series of double blind datasets,

which indicates the generalizability of the AI model. A double blind dataset is data that is collected from clinics where none of their data was used to train the AI model.

Typically performance is incorrectly quoted on a dataset that is not properly double-blinded, or sourced form a different (unseen) clinic not part of the original distribution of training data. The performance also should not just include accuracy at prediction, which can be biased if the distribution of the outcomes is biased toward viable or non-viable embryos. Therefore, it is best to quote a series of well-accepted metrics, such as sensitivity, specificity, or the confusion matrix, which contains much of the necessary information to judge efficacy. It is also useful to be able to see the distribution of AI scores themselves (i.e. how the scores are distributed in all four quarters of the confusion matrix). Other end points, such as cohort-ranking studies, are also useful, and it depends critically on how the AI is intended to be used in clinics in what the appropriate measurement is.

Q: What are your plans to conduct clinical studies that measure whether using your algorithm improves live birth rates/ TTC?

A: Our AI is aimed at looking at embryo viability, where the outcome is linked to pregnancy outcome and not live birth – many factors beyond embryo viability can result in no live birth. An increase in pregnancy outcomes will naturally increase live birth rates and decrease TTC (they are highly correlated, for obvious reasons).

Q: to all: All AI are based on fresh transferred embryos, and their implantation rate? Is there anyone who takes into account the frozen/thawed embryos and implantation rates>?

A: The Life Whisperer AI is trained on embryos that are imaged prior to freezing, and this is used as an indicator for the viability of the embryo regardless of whether it is frozen/thawed or fresh transfer. The implantation rate between the two cases is not specifically taken into account, as the AI does not directly predict implantation, since processes performed after imaging (clinical process etc) cannot be known at the time of imaging. Therefore, the AI can only be used as a predictor of the viability at Day 5 (taken as a snapshot in this case) prior to implantation, which is, of course, correlated with implantation rates.