ORIGINAL ARTICLE: ARTIFICIAL INTELLIGENCE

An interpretable machine learning model for predicting the optimal day of trigger during ovarian stimulation

Michael Fanton, Ph.D.,[a] Veronica Nutting,[a] Funmi Solano,[a] Paxton Maeder-York, M.S., M.B.A.,[a] Eduardo Hariton, M.D., M.B.A.,[b] Oleksii Barash, Ph.D.,[c] Louis Weckstein, M.D.,[c] Denny Sakkas, Ph.D.,[d] Alan B. Copperman, M.D.,[e] and Kevin Loewke, Ph.D.[a]

[a] Alife Health, Inc., Cambridge, Massachusetts; [b] Department of Obstetrics, Gynecology and Reproductive Sciences, University of California, San Francisco; [c] Reproductive Science Center, San Ramon, California; [d] Boston IVF—The Eugin Group, Waltham, Massachusetts; and [e] Reproductive Medicine Associates of New York, New York, New York

Objective: To develop an interpretable machine learning model for optimizing the day of trigger in terms of mature oocytes (MII), fertilized oocytes (2PNs), and usable blastocysts.
Design: Retrospective study.
Setting: A group of three assisted reproductive technology centers in the United States.
Patient(s): Patients undergoing autologous in vitro fertilization cycles from 2014 to 2020 (n = 30,278).
Intervention(s): None.
Main Outcome Measure(s): Average number of MII oocytes, 2PNs, and usable blastocysts.
Result(s): A set of interpretable machine learning models were developed using linear regression with follicle counts and estradiol levels. When using the model to make day-by-day predictions of trigger or continuing stimulation, possible early and late triggers were identified in 48.7% and 13.8% of cycles, respectively. After propensity score matching, patients with early triggers had on average 2.3 fewer MII oocytes, 1.8 fewer 2PNs, and 1.0 fewer usable blastocysts compared with matched patients with on-time triggers, and patients with late triggers had on average 2.7 fewer MII oocytes, 2.0 fewer 2PNs, and 0.7 fewer usable blastocysts compared with matched patients with on-time triggers.
Conclusion(s): This study demonstrates that it is possible to develop an interpretable machine learning model for optimizing the day of trigger. Using our model has the potential to improve outcomes for many in vitro fertilization patients. (Fertil Steril® 2022;■:■–■. ©2022 by American Society for Reproductive Medicine.)
Key Words: Artificial intelligence, in vitro fertilization, machine learning, ovarian stimulation, trigger

DIALOG: You can discuss this article with its authors and other readers at https://www.fertstertdialog.com/posts/34691

The goal of ovarian stimulation during in vitro fertilization (IVF) cycles is to promote multifollicular development to retrieve multiple high-quality oocytes. During ovarian stimulation, physicians make a series of decisions that are critical to the outcome of the cycle, such as which protocol to use and what starting doses of gonadotropins to prescribe. One of the most important decisions is when to give the final trigger injection to induce the final follicular maturation. Triggering too early may not allow the oocytes to reach maturity, whereas triggering too late may result in postmature oocytes as well as an increased risk of ovarian hyperstimulation syndrome. The optimal time to administer the trigger injection is a subjective decision and varies widely across practices and physicians, with limited data supporting any strict objective criteria.

Numerous studies have explored the relationship between follicle sizes and mature (MII) oocyte outcomes, reporting that follicles that are either too small or too large are less likely to yield MII oocytes (1–4). Modeling techniques have been used in retrospective studies to identify follicles of size 12–19 mm on the day

Received January 28, 2022; revised March 14, 2022; accepted April 4, 2022.
M.F. reports grant from Alife Health for the submitted work and travel support and stock options from Alife Health outside the submitted work. V.N. has nothing to disclose. F.S. has nothing to disclose. P.M.-Y. reports grant from Alife Health for the submitted work and patent and stock options from Alife Health outside the submitted work. E.H. is Medical Advisor for Alife Health and reports stock options from Alife Health. O.B. is on the Scientific Advisory Board and reports stock options from Alife Health outside the submitted work. L.W. has nothing to disclose. D.S. reports consulting fees and stock options form Alife Health. A.B.C. reports stock options for Sema4 (Chief Medical Officer) and Progyny (Medical Director). K.L reports grant from Alife Health for the submitted work and stock options from Alife Health outside the submitted work.
Supported by Alife Health.
Reprint requests: Kevin Loewke, Ph.D., 3717 Buchanan Street, Suite 400, San Francisco, California 94123 (E-mail: kloewke@alifehealth.com).

Fertility and Sterility® Vol. ■, No. ■, ■ 2022 0015-0282

https://doi.org/10.1016/j.fertnstert.2022.04.003

of trigger as having the highest likelihood of producing mature oocytes (5). The ideal way to determine which follicle sizes yield the highest maturation rates would be through individual follicle aspiration; however, such studies are difficult to conduct on a large scale. One study performed individual aspiration of binned follicle group sizes (>18 mm, 16–18 mm, 13–15 mm, 10–12 mm, and <10 mm) and showed that oocyte maturation rates increase with follicle size; however, the point at which follicles reach post-maturity was not established (6). Although previous studies have established that there is a high correlation between follicle sizes and mature oocyte outcomes, it remains unclear how to apply these findings for optimizing the timing of trigger for an individual patient.

In recent years, the field of assisted reproduction has recognized that artificial intelligence techniques can be used to support clinical decision-making during ovarian stimulation, especially with regard to optimizing the day of trigger. One of the earliest studies in this area used machine learning algorithms trained on follicle sizes and other parameters to predict whether a physician would continue stimulation or trigger (7). A more recent study developed causal machine learning techniques using follicle sizes, estradiol (E2) levels, and patient parameters to recommend continuing stimulation or trigger, with the goal of maximizing the number of fertilized oocytes (2PNs) and blastocysts (8). A potential limitation of these earlier studies is that they relied on black-box machine learning algorithms, which are unable to explain the basis for their recommendations. As the IVF field considers the adoption of these promising technologies, there is a question of whether physicians will trust black-box approaches or will prefer interpretable techniques that can explain their reasoning. As such, there is a need for further studies that explore the use of interpretable artificial intelligence for clinical decision support.

In this study, we present an interpretable machine learning approach for predicting the optimal day of trigger during ovarian stimulation. Our model uses a set of linear regression models to predict the number of MII oocytes retrieved if triggered "today vs tomorrow." A key advantage of this approach is that it can explain the basis for the recommendations. In addition, our approach provides accurate predictions of next-day E2 levels and MII oocyte outcomes that may be helpful to the physician and can be used for patient counseling. We demonstrate the performance of the model using real-world simulations in which the decisions of continuing stimulation or waiting are evaluated on a day-by-day basis. Our hypothesis for this study was that the use of an interpretable machine learning algorithm for optimizing the day of trigger may result in improved outcomes while keeping E2 levels within a safe range.

MATERIALS AND METHODS

Ethics Approval

This study was conducted after the research protocol was approved by the WCG Institutional Review Board (Study no. 1308073). Patient information was deidentified before analysis.

Study Design and Participants

This was a retrospective study using data collected from three different IVF clinics in the United States. Historical, deidentified electronic medical record (EMR) data were collected for IVF retrieval cycles started between 2014 and 2020. Records were filtered for autologous, noncanceled cycles. A total of 30,278 cycles were included in this study. An overview of the study design and methodology is shown in Supplemental Figure 1 (available online).

Data Preparation

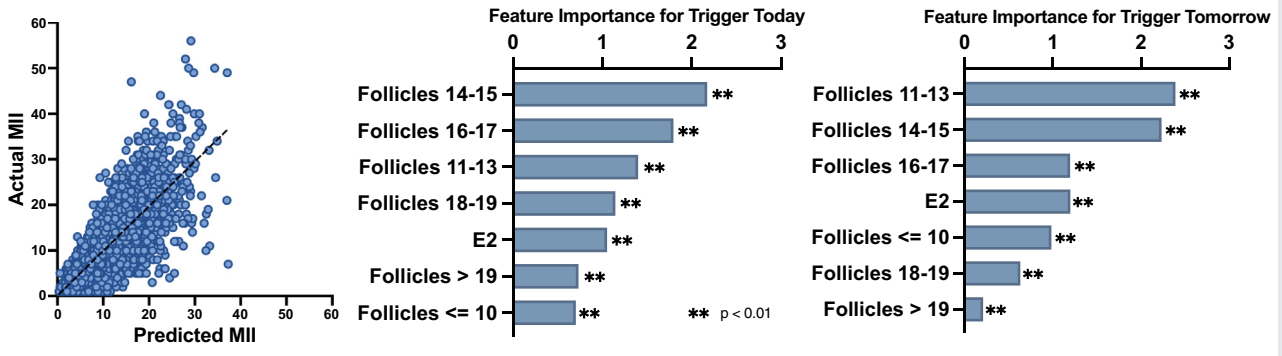
Data for training and testing the models were parsed from the EMRs. The parameters investigated included age, body mass index (BMI), number of previous IVF cycles, baseline antral follicle count (AFC), baseline anti-müllerian hormone (AMH) levels, baseline E2 levels, cycle length in days, and day-by-day measurements of follicle sizes and E2 levels from monitoring visits during ovarian stimulation. Follicle measurements were binned into 6 groups on the basis of their diameter: <11 mm, 11–13 mm, 14–15 mm, 16–17 mm, 18–19 mm, and >19 mm. The primary outcome was the number of MII oocytes retrieved, and cycles were excluded if they were missing this outcome. Additional outcomes included 2PNs and usable blastocysts (defined as the total number of transferred and frozen blastocysts). Furthermore, cycles with apparent data entry errors were excluded, such as cycles in which the number of MII oocytes exceeded the number of oocytes retrieved. Data were split by patient ID into train (70%), validation (10%), and test (20%) data sets stratified by three sites. Performance in this article is reported on the test data set. Cycles from all stimulation protocols were included in the study. After data preparation, there were 2,555 cycles from site 1, 13,051 cycles from site 2, and 14,672 cycles from site 3.

Modeling Strategy

Models were developed to enable the prediction of MII outcomes if a patient is triggered in the current day (today) compared with the next day (tomorrow), with the goal of having model interpretability. To predict the number of MII oocytes retrieved if triggered today or tomorrow, linear regression models were developed using follicle counts and E2 levels measured on the day of trigger and one day before the day of trigger, respectively. Finally, an E2 forecasting model was developed to predict next-day E2 levels using follicle counts and E2 levels measured 1 day earlier. Together, the combination of these models allowed a comparison of MII outcomes if triggered today vs. tomorrow.

Linear regression model development. Linear regression models were developed to predict the outcome of MII oocytes and next-day E2 levels using the candidate parameters previously mentioned. Approximately 83% of patient cycles had follicle and E2 measurements on the day of trigger, which were used to develop the same-day (trigger today) MII oocyte prediction model. Approximately 57% of patient cycles had follicle and E2 measurements on the day before trigger, which were used to develop the next-day (trigger tomorrow) MII

FIGURE 1



Summary of linear regression models for predicting same-day mature (MII) oocytes (trigger today) and next-day MII oocytes (trigger tomorrow). *Left:* Comparison of predicted vs actual MII oocytes. *Middle:* Order of feature importance for same-day MII model. *Right:* Order of feature importance for next-day MII model. The feature importance was taken directly from the standardized coefficients of the linear regression model.

Fanton. Machine learning for trigger optimization. Fertil Steril 2022.

oocyte prediction and E2 prediction models. All input parameters were standardized by subtracting the mean and dividing by the standard deviation of the training data set. Recursive feature elimination was used to identify the most significant parameters. Final models using only the most significant features were trained on the training data set and tuned on the validation data set, and performance was evaluated on the test data set by calculating the mean absolute error (MAE) and R^2 between the predicted number of MII oocytes and actual number of MII oocytes.

Follicle imputation. To increase the reliability of follicle measurements, a simple follicle imputation algorithm was implemented for each patient record. This algorithm employed a growth constraint on the binned follicle measurements starting on the second monitoring visit to ensure that each observed follicle either grew or remained the same size as the cycle continued.

Trigger-day recommendation algorithm. A trigger-day recommendation algorithm was developed using the set of linear regression models. For each patient in the test data set, the algorithm evaluated each of the monitoring visits day-by-day, starting from day 7, to simulate real-world scenarios. For each day of stimulation, the number of MII oocytes were predicted if triggered today and if triggered tomorrow, and the E2 level tomorrow was predicted (Supplemental Fig. 2). If the predicted number of MII oocytes today vs tomorrow showed an increasing trend, the algorithm recommended to continue stimulation. If the predicted number of MII oocytes today vs. tomorrow showed a decreasing trend, the algorithm recommended to trigger. In addition, if the end of the stimulation cycle was reached and the model had not yet recommended trigger, the model would recommend continuing another day only if the MII trend continued to increase and the predicted number of MII oocytes was <15 and the predicted E2 was $<5{,}000$ pg/mL.

Expected benefit from using the trigger model. Data from patients in the test set were used to calculate the expected benefit of using the trigger-day recommendation model. By comparing the model recommendations with the actual trigger days, patients were classified as having an early, on-time, or late trigger. To adjust for factors related to being triggered early or late, propensity score matching was used to match patients in the early/late group with patients in the on-time group. Propensity scores were calculated by training a logistic regression model to predict whether a patient would be triggered early/late or on-time using age, BMI, baseline AMH level, and baseline AFC, on all patients in the test set. Each patient's propensity score was taken as the predicted probability output of the logistic regression model. Each patient in the early/late group was matched 1:1 with the patient in the on-time group with the closest propensity score. Patients with early or late triggers were then compared with matched patients with on-time triggers in terms of average MII oocytes, 2PNs, and usable blastocysts to calculate the expected benefit of using the model. A small subset of cycles, which included MII outcomes but did not measure usable blastocysts (e.g., cleavage-stage embryo transfers or oocyte freezing cycles), were included for training and testing the MII prediction models, but they were excluded from the expected benefit analysis.

RESULTS

Supplemental Table 1 summarizes the patient demographics and cycle information for the cycles included in the study. The average day of trigger in our data set was 11.8± 1.9, and the average number of monitoring visits per cycle was 4.5± 1.4. The linear regression model for predicting MII outcomes on the day of trigger had a MAE of 2.87 oocytes and an R^2 of 0.64 on the test data set, and the model for predicting the next-day MII outcomes had an MAE of 3.02 oocytes and an R^2 of 0.62 on the test data set. The next-day E2 levels were predicted with a MAE of 274 pg/mL and R^2 of 0.88. Implementation of the follicle imputation algorithm improved the MAE by 0.09 oocytes and R^2 by 0.02. We also tried using generalized linear models with Poisson or negative binomial

distributions, because they can be appropriate for data in which the response variable contains positive integers. However, in practice, we found that regular linear regression performed the best (Supplemental Table 2).

Feature Importance

The most important feature for predicting the outcome of MII oocytes on the day of trigger was follicles of size 14–15 mm, followed by follicles of size 16–17 mm (Fig. 1). The least important feature was large follicles of size >19 mm. For predicting the outcome of the next-day MII oocytes, follicles of size 11–13 mm were of greatest importance, whereas follicles of size >19 mm remained the least important. Precycle parameters such as age, BMI, AMH level, and AFC were not significant in a multivariate model that included follicle counts and E2 levels. To ensure data quality and to evaluate follicle measurement noise, a separate model was developed to predict MII oocytes outcomes using follicle counts from the left and right ovaries separately. This model showed that the left and right follicle counts had similar coefficients for all bins, with the exception of the smallest bin (≤10 mm), and followed the same feature importance trends as the original model that adds together the follicles from both ovaries (Fig. 2). Unstandardized model coefficients are shown in Supplemental Table 3, representing how a unit change in each predictor variable would change the predicted MII outcome.

Early and Late Triggers Result in Fewer MII Oocytes and 2PNs Compared with an Optimal Trigger

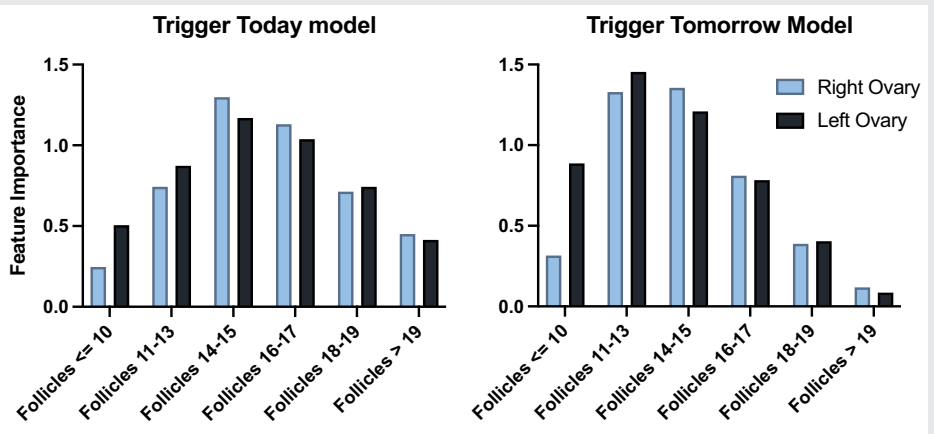
In the test data set, possible early and late triggers were identified in 48.7% and 13.8% of cycles, respectively, by comparing the actual day of trigger with what the model would have recommended (Fig. 3). Patient parameters (age, BMI, baseline AMH level, and baseline AFC) were different across early, on-time, and late trigger groups, indicating that propensity matching was appropriate. Across all test data, patients with early triggers had on average 5.3 fewer MII oocytes, 4.0 fewer 2PNs, and 2.0 fewer usable blastocysts compared with patients with on-time triggers. Patients with late triggers had on average 7.9 fewer MII oocytes, 6.0 fewer 2PNs, and 2.9 fewer usable blastocysts compared with patients with on-time triggers. After propensity score matching, patients with early triggers had on average 2.3 fewer MII oocytes, 1.8 fewer 2PNs, and 1.0 fewer usable blastocysts, and patients with late triggers had on average 2.7 fewer MII oocytes, 2.0 fewer 2PNs, and 0.7 fewer usable blastocysts compared with matched patients with on-time triggers (Table 1). After propensity matching, the patient parameters between groups were not statistically different.

A subanalysis was performed for patients who did not reach the threshold of 15 predicted MII oocytes (85% of all cycles). In this group, possible early and late triggers were identified in 59.3% and 16.8% of cycles, respectively. Patients with early triggers had on average 1.0 fewer MII oocytes, 0.6 fewer 2PNs, and 0.4 fewer usable blastocysts, and patients with late triggers had on average 3.6 fewer MII oocytes, 2.7 fewer 2PNs, and 1.3 fewer usable blastocysts compared with patients with on-time triggers.

DISCUSSION

This study is one of the first to develop an interpretable machine learning model for optimizing the day of trigger during ovarian stimulation. Our results show that over half of all cycles had possible early or late triggers on the basis of retrospective analysis. After propensity score matching, patients with early triggers had on average 2.3 fewer MII oocytes, 1.8 fewer 2PNs, and 1.0 fewer usable blastocysts, and patients

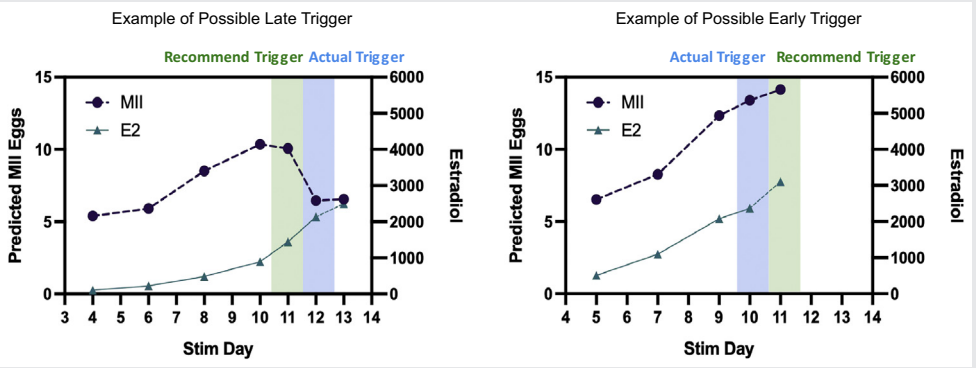
FIGURE 2



Summary of linear regression model standardized coefficients when separating follicle bins by left and right ovaries. Coefficients showed consistency between ovaries, and overall feature importance was similar to that of the original model, which adds together follicles from both ovaries.

Fanton. Machine learning for trigger optimization. Fertil Steril 2022.

FIGURE 3



Example results in which the day of recommended trigger was different from that of the actual trigger. *Left:* Example of a possible late trigger in which the model recommends trigger on day 11 but the actual trigger was on day 12. *Right:* Example of a possible early trigger, in which the model recommends trigger on day 11 but the actual trigger was on day 10. Note that in testing our model, the decision of trigger or continue stimulation was evaluated on a day-by-day basis starting from day 7 of stimulation. Only the final decision is shown here for simplicity.

Fanton. Machine learning for trigger optimization. Fertil Steril 2022.

with late triggers had on average 2.7 fewer MII oocytes, 2.0 fewer 2PNs, and 0.7 fewer usable blastocysts compared with matched patients with on-time triggers. These results indicate that significant improvements in outcomes could potentially be achieved for over half of all ovarian stimulation cycles by following the recommendations of our model.

This study places an emphasis on using interpretable machine learning techniques. Previous studies on optimizing the day of trigger have used a machine learning model with bagged decision trees, consisting of an ensemble of 30 different models (8). Such an approach has the advantage of capturing more complex and nonlinear relationships between the input parameters. This important work was one of the first to demonstrate that it is possible to use machine learning for trigger optimization. However, black-box models are inherently difficult to interpret and could have unseen problems such as overfitting or spurious correlations. The results reported in an earlier study (8) showed that outcomes could be improved by 1.43 more 2PNs and 0.57 more usable blastocysts per cycle on average using their model. Our results also show comparable improvements in 2PNs and usable blastocysts and are important for two reasons. First, our work confirms the previously reported results but across multiple different clinics and with a much larger sample size. Second, our work shows that a linear, interpretable model may provide performances similar to those of more complicated black-box models. This is not surprising, given that the inputs to these models consist of a small number of parameters. Indeed, when applying more complicated models such as random forest and XGBoost regressors to our data set to predict the outcome of MII oocytes, we found that these more complex models did not significantly improve the accuracy of prediction, despite efforts to optimize hyperparameters (Supplemental Table 2). For nonlinear models, relative feature importance can be retrospectively inferred using various post hoc analyses; however, these analyses do not directly explain how features are used to generate predicted outputs, and therefore the models remain black-box. As the field of assisted reproduction investigates the use of machine learning technologies, model interpretability will likely be an important part of achieving clinical trust and adoption (9).

The standardized model coefficients indicate that follicles of 14–15 mm and 16–17 mm in diameter were most important for predicting the outcome of MII oocytes if triggered today, whereas follicles of $\leq$10 mm and >19 mm were the least important (Fig. 1). This result is in line with the accepted notion that very small follicles are less likely to yield a mature oocyte (5, 6) and supports the idea that larger follicles may degenerate or produce postmature oocytes (10). Our model provides greater fidelity than the standard practice of monitoring lead follicle size and could provide physicians with more detailed temporal information to make an informed trigger decision. For example, if a decrease in the predicted number of MII oocytes were to be observed on consecutive days, the model coefficients would suggest that the lead follicles are growing too big, which can be verified by the physician. Conversely, an increase in the predicted number of MII oocytes could suggest that the smaller follicles are continuing to grow into the optimal size range, which also can be confirmed.

Unstandardized model coefficients allow for direct interpretability on how one additional follicle of any given size changes the predicted MII outcome. For example, as shown in Supplemental Table 3, a single follicle of size 11–13 mm contributes 0.52 oocytes for the trigger today model and 0.71 oocytes for the trigger tomorrow model. For a patient with 10 follicles of size 11–13 mm, this represents an additional 1.9 MII oocytes that could be obtained from this follicle group by waiting one additional day before trigger. Of course, this must be balanced against the potential loss of MII oocytes from other follicles that are growing too large. As shown in Supplemental Table 3, a single follicle of size 16–17 mm

TABLE 1

Comparison of patient parameters and laboratory outcomes for patients in the test data set with on-time triggers, early triggers, and late triggers, as determined by the model

Patient parameters	Early triggers	On-time triggers (matched with early)	Late triggers	On-time triggers (matched with late)
Sample size, n (%)	2,416 (48.7)	2,416	685 (13.8)	685
Age (y)	37.3	37.3	37.3	38.1
BMI (kg/m^2)	25.3	25.3	26.7	27.9
Baseline AMH (ng/mL)	1.95	2.04	1.34	1.42
Baseline AFC	10.6	10.2	8.5	8.7
No. of MII oocytes	8.4	10.7	5.8	8.5
No. of 2PNs	6.5	8.3	4.4	6.4
No. of usable blastocysts	3.2	4.2	2.4	3.1

Note: Patients with early and late triggers were propensity matched to patients with on-time triggers. Differences between on-time and early/late triggers were all statistically significant (P value $<.01$) for MII oocytes, 2PNs, and usable blastocysts. AFC = antral follicle count; AMH = antimüllerian hormone; BMI = body mass index; MII = mature oocyte; 2PNs = fertilized oocyte.

Fanton. Machine learning for trigger optimization. Fertil Steril 2022.

contributes 0.69 oocytes for the trigger today model and 0.57 oocytes for the trigger tomorrow model. For a patient with 10 follicles of size 16–17 mm, this represents 1.2 fewer MII oocytes that would be obtained from this follicle group by waiting another day. The full linear regression model weighs the relative contributions of each follicle group and provides an objective assessment of the potential benefit or harm of waiting another day before the trigger.

In developing our model, we explored different outcomes such as oocytes retrieved, MII oocytes, 2PNs, and usable blastocysts. Although all of these outcomes are important, the goal of a successful IVF cycle is of course a healthy live birth. It was not possible with our data to calculate accurate cumulative live birth rates, but our estimates showed that the oocytes retrieved, MII oocytes, and 2PNs all had a positive correlation with live birth outcomes, agreeing with previous studies (11). However, there may be a plateau for higher responders, which has been suggested in other studies (12, 13). We decided to use MII oocytes as our primary outcome, because the number of mature oocytes is a direct outcome of ovarian stimulation, whereas other outcomes such as 2PNs and blastocysts depend on the quality of sperm and laboratory procedures. We note that most of our data came from intracytoplasmic sperm injection cycles in which MII oocytes can be most accurately counted. However, we also included conventional IVF cycles if the EMR had an entry for MII oocytes, as we found that excluding conventional IVF cycles had no meaningful impact on our model performance, improving model accuracy by less than 1%.

In evaluating our model performance, the recommendation would default to trigger if the end of the cycle had been reached and the predicted number of MII oocytes had surpassed ≥ 15 oocytes or the predicted E2 level had surpassed $\geq 5{,}000$ pg/mL. These cutoffs would of course vary between physicians and clinics, but the purpose of using them was to demonstrate that we can optimize outcomes within a reasonable range, and not simply push the high responders to yield more oocytes, at the cost of increased hyperstimulation risk. Furthermore, among only patients who did not exceed the 15 MII oocyte cutoff, we showed improved outcomes for on-time patients compared with late and early patients, demonstrating that optimizing the trigger timing could improve outcomes among patients who are not high responders.

Our model to predict the outcome of MII oocytes relies on the accurate measurement of follicle sizes. However, it is known that follicle measurements can be subject to intra- and interobserver variability (14–16). In addition, as the number of follicles increases, our data suggest that small follicles may not always be counted. The reasons for this are not clear, but it could be because more importance is placed on the larger follicles that are expected to yield a mature oocyte or that follicles that are unlikely to grow enough are ignored closer to the time of trigger. Other possibilities include vanishing follicles, which is a rarely observed phenomenon in women with advancing age (17). In general, however, we expect that follicles should not disappear during the ovarian stimulation process. For this reason, we implemented a follicle imputation algorithm to make sure that the total follicle count always stayed the same or increased. This imputation helped achieve a small improvement in the performance of our model. Linear regression models trained on follicles separated by ovary show that the right and left ovary bins had very similar coefficients for follicle bin sizes, with the exception of the smallest bin (≤ 10 mm) that had a larger coefficient for the left ovary than that for the right ovary. This observation is likely a result of the aforementioned inconsistencies in small follicle measurements across sites. Despite this, the small follicle bin for our original model, which adds together the follicles from both ovaries, had a strong predictive value in our models and was therefore kept as an input feature. We believe that successful clinical use of our model in the future will depend on accurate follicle counting at each monitoring visit during ovarian stimulation.

This study is not without limitations; the primary limitation is its retrospective nature. We did not differentiate between different trigger medications or types of protocols, which should be explored further in future work. Some cycles in our data set had incomplete or missing data. For example,

in some cycles, there were no ultrasound measurements taken on the day of trigger. These cycles were therefore excluded from our analysis, which could have introduced a sampling bias. It is possible that our results and feature importance could be biased by the trigger practice of the clinics, for example, if patients are always triggered when two or more follicles reach 18 mm in size. However, the possibility for bias should be reduced by the fact that our training data comes from three separate clinics with varying practices on when to trigger patients. In addition, our data include patients pushed further than a clinic's general guidelines as well as those triggered earlier on the basis of E2 levels. Additionally, although we implemented an E2 level threshold as a surrogate marker for the risk of ovarian hyperstimulation syndrome, we did not have data on which patients had actually been hyperstimulated. Similarly, our models were not able to incorporate risks of other adverse clinical events, such as premature ovulation, given that our data set included only completed cycles. Expanding our data set in future work will allow for the ability to train models to recognize patients at risk for these complications. Finally, our model did not take into account a patient's previous stimulation results, as our analysis did not show that outcomes from a patient's first stimulation could further improve predictions of outcomes in their second stimulation. However, this should also be further investigated in future work.

CONCLUSION

This study is one of the first to develop an interpretable machine learning model for optimizing the day of trigger during ovarian stimulation. Our results indicate that an interpretable machine learning model can potentially improve outcomes in a considerable number of patients. Future work will focus on continuing to increase the size and diversity of our training data set and performing prospective validation studies to show improved patient outcomes with the use of our model.

Acknowledgments: The authors thank Dmitry Gounko, Brianna Amaral, and Daniel Duvall for their help with data collection. These important data made this study possible. The authors also thank the team members of Alife Health for their helpful discussions and review of the manuscript.

DIALOG: You can discuss this article with its authors and other readers at https://www.fertstertdialog.com/posts/34691

REFERENCES

1. Dubey AK, Wang HA, Duffy P, Penzias AS. The correlation between follicular measurements, oocyte morphology, and fertilization rates in an in vitro fertilization program. Fertil Steril 1995;64:787–90.
2. Salha O, Nugent D, Dada T, Kaufmann S, Levett S, Jenner L, et al. The relationship between follicular fluid aspirate volume and oocyte maturity in in-vitro fertilization cycles. Hum Reprod 1998;13:1901–6.
3. Nogueira D, Friedler S, Schachter M, Raziel A, Ron-El R, Smitz J. Oocyte maturity and preimplantation development in relation to follicle diameter in gonadotropin-releasing hormone agonist or antagonist treatments. Fertil Steril 2006;85:578–83.
4. Mehri S, Levi Setti PE, Greco K, Sakkas D, Martinez G, Patrizio P. Correlation between follicular diameters and flushing versus no flushing on oocyte maturity, fertilization rate and embryo quality. J Assist Reprod Genet 2014;31:73–7.
5. Abbara A, Vuong LN, Ho VNA, Clarke SA, Jeffers L, Comninos AN, et al. Follicle size on day of trigger most likely to yield a mature oocyte. Front Endocrinol (Lausanne) 2018;9:193.
6. Rosen MP, Shen S, Dobson AT, Rinaudo PF, McCulloch CE, Cedars MI. A quantitative assessment of follicle size on oocyte developmental competence. Fertil Steril 2008;90:684–90.
7. Letterie G, Mac Donald A. Artificial intelligence in in vitro fertilization: a computer decision support system for day-to-day management of ovarian stimulation during in vitro fertilization. Fertil Steril 2020;114:1026–31.
8. Hariton E, Chi EA, Chi G, Morris JR, Braatz J, Rajpurkar P, et al. A machine learning algorithm can optimize the day of trigger to improve in vitro fertilization outcomes. Fertil Steril 2021;116:1227–35.
9. Wang R, Pan W, Jin L, Li Y, Geng Y, Gao C, et al. Artificial intelligence in reproductive medicine. Reproduction 2019;158:R139–54.
10. Revelli A, Martiny G, Delle Piane L, Benedetto C, Rinaudo P, Tur-Kaspa I. A critical review of bi-dimensional and three-dimensional ultrasound techniques to monitor follicle growth: do they help improving IVF outcome? Reprod Biol Endocrinol 2014;12:107.
11. Hariton E, Kim K, Mumford SL, Palmor M, Bortoletto P, Cardozo ER, et al. Total number of oocytes and zygotes are predictive of live birth pregnancy in fresh donor oocyte in vitro fertilization cycles. Fertil Steril 2017;108:262–8.
12. Polyzos NP, Drakopoulos P, Parra J, Pellicer A, Santos-Ribeiro S, Tournaye H, et al. Cumulative live birth rates according to the number of oocytes retrieved after the first ovarian stimulation for in vitro fertilization/intracytoplasmic sperm injection: a multicenter multinational analysis including ~15,000 women. Fertil Steril 2018;110:661–70.e1.
13. Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. Hum Reprod 2011;26:1768–74.
14. Eissa MK, Hudson K, Docker MF, Sawers RS, Newton JR. Ultrasound follicle diameter measurement: an assessment of interobserver and intraobserver variation. Fertil Steril 1985;44:751–4.
15. Vandekerckhove F, Bracke V, De Sutter P. The value of automated follicle volume measurements in IVF/ICSI. Front Surg 2014;1:18.
16. Lujan ME, Chizen DR, Peppin AK, Kriegler S, Leswick DA, Bloski TG, et al. Improving inter-observer variability in the evaluation of ultrasonographic features of polycystic ovaries. Reprod Biol Endocrinol 2008;6:30.
17. Dozortsev DI, Diamond MP. Two peas from the same pod: vanishing follicles and postmature oocytes. Fertil Steril 2022;117:40–1.