Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)
© 2021 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA210141

Limits of Conventional Machine Learning Methods to Predict Pregnancy and Multiple Pregnancy After Embryo Transfer

Núria CORREA[a,1], Rita VASSENA[a], Jesús CERQUIDES[b], Josep Lluís ARCOS[b]
[a] *Clinica Eugin-Eugin Group*
[b] *IIIA-CSIC*

Abstract. When training models to learn the relationship between two or more variables, we expect to see previously demonstrated knowledge about that relationship reflected in the resulting estimators. For some domains, such as healthcare, it is imperative for actual implementation of those models that their predictions respect this knowledge. In this study we focus on Assisted Reproduction Technology (ART), the subspecialty of gynecology occupied with treating human infertility, and where the goal of any treatment is the delivery of a healthy newborn. A common ART treatment is In vitro Fertilization (IVF), where embryos are generated in vitro from collected sperm and oocytes, and transferred to the uterus of the patient after selecting those most likely to give rise to a healthy pregnancy. IVF has an approximate 30% successes rate per cycle; to palliate for this low success rate, a common practice so far has been to transfer two embryos simultaneously, aiming to increase the chances of a favorable outcome. While increasing overall live birth rates, this method has also led to an alarmingly high rate of twin and triplet births, associated with four times higher risk of perinatal mortality and increased obstetric complications. Our objective is to predict the chances of both pregnancy (P) and multiple pregnancy (MP) following either single embryo transfer (SET) or double embryo transfer (DET), and in so facilitating an informed decision on how many embryos to transfer. From existing literature, it is known that: (1) it is not possible for the chances of both P and MP to be decreased by increasing the number of embryos; (2) MP chances cannot be higher than P; and (3) chances of pregnancy are highly correlated with age, embryo stage, and quality. With a dataset generated from an existing observational study, we trained several state-of-the-art classifiers to predict P and MP given SET and DET. Analyzing the results, all classifiers achieved promising AUC scores. However, Random Forest and Gradient Boosting predicted negative chance differences in many instances when increasing the number of embryos infringing the first constraint. Logistic Regression predicted always positive differences, but in some instances it infringes the second constraint, predicting higher chances of MP than of P. Moreover, it showed little to no variation across ages or embryo stages violating third constraint. Conventional Machine Learning models struggle to reflect the real-world outcomes when using DET versus SET in specific patients. More informative variables could help, but it is already worrisome that variables as important as age and embryo stage do not result already in any variation, and that when models do show variation, in many cases they predicted decreasing chances of success with more embryos. We conclude that new and different approaches are needed to correctly model this scenario and, likely, many others resembling this one.

Keywords. Machine Learning, Classification, Healthcare

[1] Núria Correa, R+D Department, Clínica Eugin, Balmes 236, 08006 Barcelona, Spain; E-mail: ncorrea@eugin.es.

1. Introduction

With the rising general popularity of Artificial Intelligence (AI) and, specifically, Machine Learning (ML), several fields have jumped on the bandwagon of applying them to different processes. One such field is healthcare, where many high stakes and fast decisions must be made. As high dimensional data registers are frequently available, it stands to reason that those could be learned from using ML. In many healthcare scenarios a heavy research background already exists, providing a high amount of evidence-based knowledge. In this context, it is expected that previously demonstrated data relations are picked up by trained models and their predictions heed them. Additionally, to ensure user confidence, the explainability of ML models is of paramount importance. Explainability also needs to be coherent with previously demonstrated knowledge. In other words, expectations on how the models will work are set by preceding research, and failure to comply with them can diminish the confidence of the users in the models' predictions.

This is the situation of our subject of interest: Assisted Reproductive Technologies (ART), a subspecialty of gynecology that is preoccupied with the instrumental treatment of human infertility and whose main goal is the delivery of a healthy newborn. In order to achieve this objective, different techniques have been developed and are applied depending on the necessities of the patient. A common kind of ART treatment is In Vitro Fertilization (IVF), where oocytes and sperm are combined in vitro to generate embryos. After selecting those expected to have better chances of giving rise to a healthy pregnancy, the embryos are transferred to the uterus of the patient. IVF provides an approximated 30% pregnancy rate per treatment, which leads to about 20% delivery rate [1] . These rates can undoubtedly be frustrating for both professionals and patients. To mitigate low success rates, the transfer of two embryos simultaneously to the uterus has been proposed. This certainly increases the chances of achieving a pregnancy versus Single Embryo Transfer (SET) [2] ; Double Embryo Transfer (DET) now represents 54.5% of all embryo transfers. Unfortunately, the increase in success comes with an increased obstetrical risk, reflected by the troublingly high 17% of twin births DET. Measured against singleton births, twin births have a four times higher risk of perinatal mortality. Twin pregnancies are also associated with an increased risk of obstetric complications, higher rates of miscarriage, pregnancy-induced hypertension, gestational diabetes, premature labor and abnormal delivery compared to singleton pregnancies [3] . As a consequence, a twin pregnancy is an undesired outcome of ART cycles.

Nevertheless, the rate of DET remains high; why is this? The issue is indeed complex. As stated before, Randomized Controlled Trials (RCTs) have consistently shown that SET provides lower pregnancy rates than DET, but they do so with the bonus of a much lower twin rate. Literature also indicates that the cumulative pregnancy rate between repeated SET and a single round of DET is similar, but there is a much lower twin rate in patients that get SET+SET vs. DET [2]. This would, from a strictly clinical point view, lead to an easy solution, which would be to always use repeated SET. But, as stated before, the issue is not that straightforward.

On the one hand, we should acknowledge that the embryos available to a woman for transfer are not always of high morphological quality, and having worse morphology is an indicator of worse development potential and higher aneuploidy rates [4]. In these cases, DET is used as a strategy to allow for higher pregnancy rates in bad prognosis treatments, assuming that the risk of multiple pregnancy should be lower as one of the two embryos transferred has low chances to implant. Further, embryo stage may influence the outcome, as there is moderate quality evidence that blastocyst stage embryos (at day 5 or 6 after fertilization) have better chances of pregnancy versus cleavage stage embryos (at day 2 or 3 after fertilization) [5]. Also, regardless of embryo quality and stage, the specifics of every case modulate the chances of pregnancy as does for example the age of the oocyte [6] and its origin (donor or own oocytes), the integrity of the uterine environment and shape, the reproductive history of the couple or single patient, the parameters and origin (donor or partner) of the semen used to fertilize the oocytes, etc. On a day-to-day basis, all this information is processed by the clinical experts in order to make a professional recommendation based on literature and hands-on experience on the adequate number of embryos to be transferred in order to achieve the highest possible live birth rates with the lowest possible multiple pregnancy rate.

On the other hand, patients are paramount in these processes, as they are the ones going through the treatment with the very emotionally charged goal of being able to give birth. They participate actively in making the final decision of how many embryos will get transferred, and often non-clinical factors weight in their decision. Some of those factors include their psychological state (affected by repeated treatments, urgency to get pregnant, previous interrupted pregnancies, etc.), the economic pressure of the treatments and the information that they receive and/or understand [7].

Considering all this, it is clear that the clinical objective when selecting between a SET or DET treatment for each individual patient is to get the highest pregnancy chance with the lowest twin pregnancy risk. And so, it is natural to search for methods that allow us to predict better the chance of pregnancy (P) and multiple pregnancy (MP) for patients before getting SET or DET. Here is where ML can be of help.

Then, the technical objective of this study is to train models able to predict chances of P and MP given a set of covariates that include both treatment options. Getting accurate models for these tasks would enhance professionals' confidence in aiding patients to make an informed decision. But in order for those models to be really regarded as usable in clinical practice they need to heed previously demonstrated knowledge, leading us to identify three main constraints:

1. Under stable conditions (same patient, same cohort of embryos) it is not possible for the chances of both P and MP to be decreased by increasing the number of embryos transferred.

2. Under stable conditions MP chances cannot be higher than P.

3. Chances of P and MP are highly correlated with age, embryo stage, and quality.

To properly test the performance of conventional ML models not only standard measures as AUC need to be analyzed, but also compliance with all three constraints needs to be examined.

Few studies have been carried out in this regard, but the ones that did give us some interesting insight. A very thorough report on the theme performed [8] recounts construction of P and MP models using first UK national reports with AUC 0.60 for the first model and 0.66 for the second, and then information from multiple private centers with more predictor variables and slightly better AUC scores. It also reviews an approach modeling separately the uterus component and the embryo component. Another study creates only an MP model for patients that got DET [9]. Lastly, another interesting study created independent models: one for P and MP on DET cycles, and another one for P on SET cycles, getting AUCs between 0.64 and 0.75 [10]. Interestingly, this last model has been also tested on patients and has helped to significantly reduce the incidence of MP. All of these studies are promising, but do not check for the first two constraints which we find are certainly critical.

2. Materials and methods

There are multiple public and published sources that report results on pregnancy and multiple pregnancy with both SET and DET [8 , 11]. All these populational studies are coherent between them but offer only summarized sample statistics, and no granular patient level datasets are publicly available. In this study, to ensure reproducibility, we focused on the data from the observational study by Aldemir et al. (2020) [11], taken as a guiding example to synthetically generate a dataset. In their study where 2298 patients were included three groups were compared: those who got DET with good quality embryos (GQEs), DET with mixed quality embryos (MQEs), and SET with good quality embryos. For those three groups several variables were gathered, including age, embryo stage, pregnancy and multiple pregnancy.

The replicated dataset was carefully constructed. Maternal age was simulated for every group using mean and standard deviation reported by the observational study to randomly sample from a normal distribution, resulting in 33.28 ± 4.1 for the first group, 34.4 ± 3.8 for the second and 29.2 ± 4.1 for the third. Individual outcomes of P and MP per group and embryo stage were sampled randomly from reported results using a Bernoulli distribution. The resulting proportions, shown in Table 1 and 2, had less than a 5% deviation compared to the original study results. Further, strict restrictions were put in place in order to avoid inconsistencies on our artificial dataset, such as cases with positive MP results but a negative P result.

Table 1. Proportions of pregnancy and multiple pregnancy instances by group in patients who got transferred embryos at the cleavage stage

	DET with GQEs **n=324**	**DET with MQEs** **n=127**	**SET with GQE** **n=887**
P	41.05	35.43	29.99
MP	23.46	9.45	3.16

Table 2. Proportions of pregnancy and multiple pregnancy instances by group in patients who got transferred embryos at the blastocyst stage

	DET with GQEs **n=174**	**DET with MQEs** **n=52**	**SET with GQE** **n=734**
P	56.32	23.08	43.46
MP	32.76	25.00	2.32

Three common ML classifiers were selected to be trained on our resulting database. Those classifiers were Logistic Regression (LR), Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC). 80% of the synthetic database was used to train them and the other 20% was reserved for testing purposes. Average AUC and accuracy scores were obtained by cross validating 10 times over the training dataset.

As not only conventional scores are important in this kind of scenarios, the predicted outcomes on the test portion were analyzed to assess compliance of the 3 stated constraints. In order to do that all patients (1) got predicted probabilities of P and MP with SET and DET separately to detect any negative "effects"; (2) got predicted probabilities of P and MP to detect cases with higher MP chances than those of P; and (3) both P and MP predicted chances were examined for its relations with maternal age and embryo stage and quality.

3. Results

After analyzing common scores as AUC and Accuracy, LR and GBC seem to be the ones that fare better at predicting both outcomes, with LR being slightly better at AUC and GBC at accuracy (see Table 3). Regarding the mean expected effect of using DET versus SET for every specific patient, all estimators get close to values described in literature regarding P, which fall between 12% and 23% increased chances [2]. This is not the case of MP, where multiple RCTs pooled suggest an increase between 11% and 13%. RFC and GBC are slightly over those values, and LR is very clearly out of the described range.

Table 3. Results of the divided by type of model (Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier) and outcome (Pregnancy and Multiple Pregnancy).

	AUC	Accuracy	Mean effect	Constraint 1	Constraint 2	Constraint 3
LR-P	0.58	0.55	0.14	Yes	No	Partial
LR-MP	0.78	0.75	0.55	Yes	-	No
RFC-P	0.52	0.54	0.17	No	No	No
RFC-MP	0.71	0.86	0.24	No	-	No
GBC-P	0.56	0.62	0.12	No	No	Partial
GBC-MP	0.77	0.91	0.21	No	-	Partial

When considering the first constraint (under the same conditions increasing the number of embryos cannot decrease the success chances), only LR complies fully with it. RFC and GBC both show multiple instances where their predictions estimate a decrease in chances in DET vs SET in the same patient, as shown for example in Figure 1.

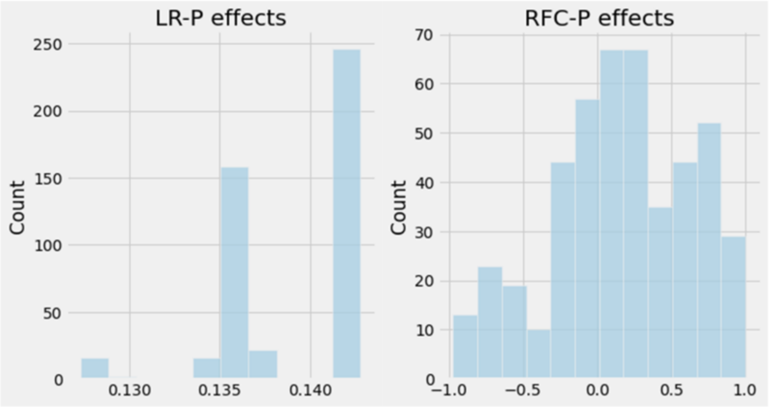

Figure 1. Probability differences between predictions on the same patients with SET and DET in the models Logistic Regression (left) and Random Forest Classifier (right) trained to predict pregnancy outcomes.

Looking upon the second constraint we found no compliance across all models studied, with GBC being the one with the least instances where the constraint was infringed (see Figure 2).

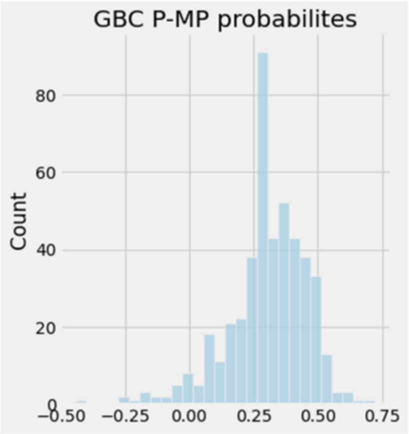

Figure 2. Distribution of the differences in predicted probabilities for pregnancy and multiple pregnancy using Gradient Boosting Classifiers.

Lastly, for the third constraint both LR and GBC comply only partially and RFC does not comply with it. It is referred as partial as with both models age seems to add little to no variation in predicted P when embryo stage does, as shown in Figure 3. Predicting MP though, GBC seems to show some variation across ages, but LR does not comply at all.

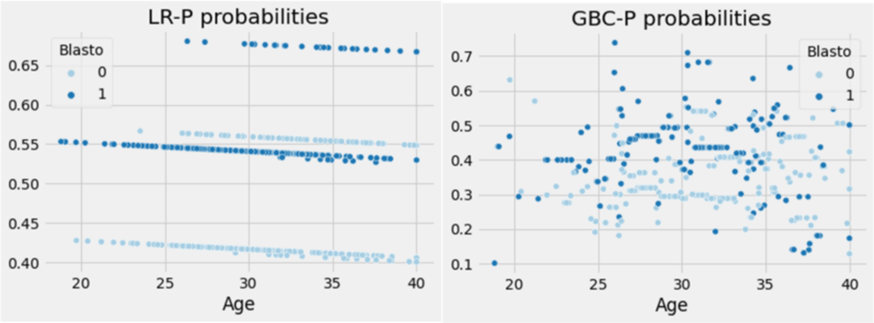

Figure 3. Logistic Regression and GBC pregnancy predicted probabilities plotted against maternal age and colored by embryo stage (blastocyst yes or no).

4. Discussion

We show here that the three conventional ML models tested are not entirely suitable for the task at hand, even if AUC scores remain close to those obtained in literature. The highest performing algorithm seems to be LR but even so, it only complies with one and a half of our presented constraints. If presented to a field expert for clinical practice, evidently it would be regarded as unfaithful and thus unusable. In any ML implementation related to healthcare, not only a model needs to be accurate, but it also needs to convince the professional about its reliability, and that means being consistent with evidence-based knowledge. Especially nowadays when AI and ML models are under public scrutiny and asked to be accountable, and that leads to be able to explain their decisions.

Concerning the first constraint, there seems to be an inbuilt bias in the dataset, where younger patients and embryos with better qualities or more advanced embryo stages tend to lead to more SET treatments. This is in agreement with previous knowledge of the field, as better prognosis is associated with a higher risk of MP, and so professionals and patients tend to prefer SET. Older patients and lower quality embryos tend to fair worse, and so with lower risks of MP, they tend to get more DET attempts. In other words, treatment is not randomized, as it is often the case in observational databases. Also, our dataset does not contain SET with embryos of worse quality, nor DET with both embryos of bad quality. This may create a confounding effect that cannot be accounted for correctly by the model. It would be interesting to identify from the literature studies with more types of embryo combinations, to understand if this may remain a concern. Unfortunately, none of the published researches check for that constraint.

As for the second constraint, one of the main problems in this approach to the matter seems to be the need to model two separate but closely related outcomes without being able to state some restrictions on how the models should predict both outcomes for the same patient. Even if treated as a multiclass problem (with outcomes failure, P, and MP) we would not be able to specify that there should never be a higher chance of MP than of P with common ML models. Looking at the available literature, a way of overriding the second constraint would be by constructing the MP model only using data of DET cycles that got a successful P, as that is what all studies do in constructing MP models. But that would drastically reduce the size of the available dataset and maybe hinder the models' performance. It also completely ignores the prediction of the probabilities of MP for SET cycles that, though they have very little chances in general of an instance of MP, could be also interesting to be able to predict.

Last but not least, the third constraint seems to be fairly complied with in previous studies on the matter where datasets include far more information, which would lead us to think that possessing a database of that characteristics would enable us to get models compliant with it.

5. Conclusions

In this work we have shown that conventional ML models, even when performing well in terms of prediction score at the population level, struggle considerably at the individual level. In doing so, they fail to comply with evidence-based derived constraints. As we stated in our motivation, in healthcare explainability is mandatory and it should always guarantee alignment with previous evidence-based knowledge. As exposed in other studies [12], failing to ensure cohesiveness can lead to diminished user confidence in the model and, in the worst-case scenario, to detrimental consequences for patients.

Focusing on our specific experiment, for the second and third constraints there seems to be possible solutions, but for the first one there seems not to be a straightforward answer. Finding a way to define beforehand the relationship between key variable treatment and outcome as monotonically ascending could take us a step closer to obtaining more realistic models. This challenge is not specific of the situation described here, rather it is somewhat endemic in the healthcare field, and so it constitutes a barrier to adopt AI solutions. Therefore, it is clear that new and different approaches to this kind of challenges would be needed.

Acknowledgments

This work was supported by Doctorat Industrial funded by Generalitat de Catalunya (DI-2019-24),by project CI-SUSTAIN funded by the Spanish Ministry of Science and Innovation (PID2019-104156GB-I00), and by intramural funding by Clinica Eugin-Eugin Group. Núria Correa is a Ph.D. Student of the doctoral program in Computer Science at the Universitat Autònoma de Barcelona.

References

[1] De Geyter C, Calhaz-Jorge C, Kupka MS, Wyns C, Mocanu E, Motrenko T, et al. ART in Europe, 2014: Results generated from European registries by ESHRE. Hum Reprod. 2018;33(9):1586–601.

[2] Kamath MS, Mascarenhas M, Kirubakaran R, Bhattacharya S. Number of embryos for transfer following in vitro fertilisation or intra-cytoplasmic sperm injection. Cochrane Database Syst Rev. 2020;2020(8).

[3] Crosignani PG, Baird DT, Barri P, Bryan E, Collins J, Diedrich K, et al. Multiple gestation pregnancy. Hum Reprod. 2000;15(8):1856–64.

[4] Hardarson T, Caisander G, Sjögren A, Hanson C, Hamberger L, Lundin K. A morphological and chromosomal study of blastocysts developing from morphologically suboptimal human pre-embryos compared with control blastocysts. Hum Reprod. 2003;18(2):399–407.

[5] Glujovsky D, Farquhar C, Quinteiro Retamar AM, Alvarez Sedo CR, Blake D. Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. Cochrane Database Syst Rev. 2016;2016(6).

[6] Grøndahl ML, Christiansen SL, Kesmodel US, Agerholm IE, Lemmen JG, Lundstrøm P, et al. Effect of women's age on embryo morphology, cleavage rate and competence - A multicenter cohort study. PLoS One. 2017;12(4):1–12.

[7] de Lacey S, Davies M, Homan G, Briggs N, Norman RJ. Factors and perceptions that influence women's decisions to have a single embryo transferred. Reprod Biomed Online. 2007;15(5):526–31.

[8] Roberts SA, McGowan L, Hirst WM, Brison DR, Vail A, Lieberman BA. Towards single embryo transfer? modelling clinical outcomes of potential treatment choices using multiple data sources: Predictive models and patient perspectives. Health Technol Assess (Rockv). 2010;14(38):1–237.

[9] Lannon BM, Choi B, Hacker MR, Dodge LE, Malizia BA, Barrett CB, et al. Predicting personalized multiple birth risks after in vitro fertilization-double embryo transfer. Fertil Steril [Internet]. 2012;98(1):69–76. Available from: http://dx.doi.org/10.1016/j.fertnstert.2012.04.011

[10] Vaegter KK, Berglund L, Tilly J, Hadziosmanovic N, Brodin T, Holte J. Construction and validation of a prediction model to minimize twin rates at preserved high live birth rates after IVF. Reprod Biomed Online [Internet]. 2019;38(1):22–9. Available from: https://doi.org/10.1016/j.rbmo.2018.09.020

[11] Aldemir O, Ozelci R, Baser E, Kaplanoglu I, Dilbaz S, Dilbaz B, et al. Impact of Transferring a Poor Quality Embryo along with a Good Quality Embryo on Pregnancy Outcomes in IVF/ICSI Cycles: a Retrospective Study. Geburtshilfe Frauenheilkd. 2020;80(8):844–50.

[12] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science (80-). 2019;366(6464):447–53.