




**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



**Universitat  
Autònoma  
de Barcelona**

PhD in Computer Science

Research line: Artificial Intelligence

**AI with care:  
Integrating machine learning  
with expert knowledge  
for In Vitro Fertilization**

PhD Student: Núria Correa Mañas

PhD Advisors: Rita Vassena, Jesús Cerquides Bueno, Josep Lluís Arcos Rosell

PhD Tutor: Josep Lluís Arcos Rosell

PhD Admission Date: 01/10/2019

Contact mail: [ncorrea@iia.csic.es](mailto:ncorrea@iia.csic.es) / [ncorrea@eugin.es](mailto:ncorrea@eugin.es)

Bellaterra (Cerdanyola del Vallès), June 7, 2023

*To my family, including the found one.*

# Abstract

After almost 45 years from the birth of Lousie Brown, the first baby born after in vitro fertilization (IVF), pregnancy rates for this treatment remain around 30%, with a 20% chance of delivery. Even if it is much better than the chances than those patients had without IVF, logically, there are constant endeavours to gain insight into the biological reality behind fertility in order to refine artificial reproduction technologies (ART).

In parallel to the technical advances achieved by ART professionals, artificial intelligence (AI) has also progressed at a remarkable pace. Its ability to deal with high dimensional databases and detect hidden data relationships has led researchers to investigate its application to healthcare. There are several processes in ART, and specifically, IVF, where AI methods are currently being applied.

In this thesis, the main focus is on the selection of the first dose of follicle-stimulating hormone (FSH) for controlled ovarian hyperstimulation (COH). COH is the first step of an IVF treatment, where the objective is to retrieve an optimal number of mature oocytes from the ovary. Its results are critical for the success of the IVF treatment. Standard clinical protocols to select the first dose of FSH are not perfect, and lead a sizable portion of patients to suboptimal results. Here, we use AI methods with historical data from past COH treatments to obtain an optimized FSH dosing policy.

Historical or observational datasets are often biased and prone to low variability due to the high adherence of clinicians to standard protocols. In this context, out-of-the-box AI methods do not have enough information to learn dosing models that improve standard practice, or even show consistency with the underlying physiological reality. Hence, the introduction of domain knowledge in the training process is key to obtain clinically robust models from observational data. To achieve this, we propose building the dosing model around the assumption that the dose-response relationship between FSH and the number of oocytes retrieved is monotonic.

Further, since insight on the performance of dosing models is generally achieved through prospective clinical intervention, we designed an ad-hoc performance score to evaluate doses (real or counterfactual) pre-clinically. This score, based on expert knowledge, can evaluate whether a dose is appropriate depending on the ground truth outcome, expressed as the number of mature oocytes

retrieved. Using this method, we were able to ascertain a statistically significant improvement versus standard clinical practice. A generalized method for similar dosing problems, called IDoser, was also tested in the FSH use case against clinical practice and a benchmark of literature, finding again significant improvement. A first approach of the application of IDoser to the selection of the number of embryos for transfer in IVF also returned positive results with potential for improvement.

Finally, AI driven solutions, especially for healthcare settings like drug dose selection, are to be handled with care, as the patients' health is at stake. Not only that, they need to gain the trust of their intended users, in this case, clinicians. Trust is gained through clinical improvement and easy-to-prove adherence to already available field knowledge. The first can be achieved through pre-clinical analysis, but especially through randomized controlled trials (RCTs) where the models are tested against standard practice. The second, as proposed in this thesis, can be achieved through interpretable implementations of domain knowledge into the creation and training of dosing models. This leads to clinically robust dosing models that achieve better pre-clinical results.

# Contents

<b>Contents</b>	<b>iv</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Thesis motivation . . . . .	1
1.2 Research questions . . . . .	3
1.3 Ethical aspects of this project . . . . .	5
1.4 PhD structure . . . . .	5
<b>2 State of the art</b>	<b>7</b>
2.1 AI in Healthcare . . . . .	7
2.1.1 Dose-response modelling . . . . .	11
2.1.2 Conclusions . . . . .	17
2.2 AI in ART . . . . .	18
2.2.1 General overview . . . . .	18
2.2.2 Dose recommendation in IVF . . . . .	30
2.3 RCT trials for AI solutions in ART . . . . .	36
2.4 Conclusions . . . . .	37

<b>3</b>	<b>FSH dosing policy optimization</b>	<b>38</b>
3.1	Background . . . . .	39
3.1.1	Patient population . . . . .	40
3.2	First iteration . . . . .	41
3.2.1	Development of a performance score . . . . .	41
3.2.2	Prediction model . . . . .	44
3.2.3	Dosing model and performance evaluation . . . . .	44
3.2.4	Results . . . . .	45
3.2.5	Lessons learned . . . . .	47
3.3	Second iteration . . . . .	47
3.3.1	Results . . . . .	50
3.4	Discussion . . . . .	52
3.5	Conclusions . . . . .	55
<b>4</b>	<b>IDoser: including field knowledge into the training of dosing models</b>	<b>56</b>
4.1	Background . . . . .	57
4.2	The Individualized Dose Improvement Problem . . . . .	58
4.3	Proposal: Individualized Doser (IDoser) . . . . .	59
4.3.1	The core model . . . . .	60
4.3.2	Loss function . . . . .	60
4.3.3	Optimization of parameters . . . . .	62
4.4	Use case . . . . .	62
4.4.1	IDoser for FSH dosing . . . . .	64
4.5	Evaluation Methodology . . . . .	65
4.5.1	Literature benchmark . . . . .	65
4.5.2	Optimization exploration . . . . .	65
4.5.3	Model comparison and statistic tests . . . . .	66
4.6	Results . . . . .	66

4.7	Discussion . . . . .	68
4.8	Conclusions . . . . .	72
<b>5</b>	<b>IDoserFSH: A non-inferiority study protocol for a multi-center randomized c trial</b>	<b>73</b>
5.1	Background . . . . .	73
5.2	Trial general details . . . . .	74
5.3	Methods: Participants, interventions and outcomes . . . . .	76
5.4	Interventions . . . . .	76
5.5	Assignment of interventions: allocation . . . . .	80
5.6	Assignment of interventions: Blinding . . . . .	81
5.7	Data collection and management . . . . .	81
5.8	Statistical methods . . . . .	82
5.9	Oversight and monitoring . . . . .	83
5.10	Conclusions . . . . .	85
<b>6</b>	<b>Limits of conventional Machine Learning methods to predict pregnancy and multiple pregnancy after embryo transfer</b>	<b>86</b>
6.1	Background . . . . .	86
6.2	Material and Methods . . . . .	88
6.3	Results . . . . .	90
6.4	Discussion . . . . .	93
6.5	Conclusions . . . . .	94
<b>7</b>	<b>Conclusions and Future Work</b>	<b>95</b>
7.1	General conclusions . . . . .	95
7.2	Contributions . . . . .	97
7.3	Publication list . . . . .	100
7.4	Future work . . . . .	102
7.4.1	IDoser method . . . . .	102



7.4.2	IDoserFSH . . . . .	102
7.4.3	IDoser for selection of number of embryos for transfer . . . . .	103
7.4.4	Final conclusions and next steps . . . . .	109
<b>Appendices</b>		<b>110</b>
<b>A Tables and figures for the FSH performance score function</b>		<b>111</b>
<b>B IDoser: Optimization exploration and statistic results</b>		<b>115</b>
B.1	Optimization exploration . . . . .	115
B.2	Statistics results . . . . .	116
<b>Bibliography</b>		<b>120</b>

# List of Figures

1.1	SHAP dependence plot for a fictional example patient. In red factors that increase the value of the number of oocytes predicted. In blue factors that decrease the prediction. Other variables present in this plot are related to ovarian reserve or patient age at the moment of treatment. . . . .	3
2.1	Results of the exercise–cholesterol study, unsegregated. Taken from Pearl, Glymour, and Jewell, 2016. . . . .	15
2.2	Results of the exercise–cholesterol study, segregated by age. Taken from Pearl, Glymour, and Jewell, 2016. . . . .	15
2.3	Directed Acyclic Graph showing a confounder ( $Z$ ) affecting both treatment ( $X$ ) and outcome( $Y$ ). . . . .	16
2.4	Graphical summary of the steps of the IVF treatment. Customized from the original template design published in canva.com by @martaborreguero. . . . .	22
3.1	Outcome ranges expressed in number of oocytes considered non-desirable, sub-optimal and optimal. . . . .	40
3.2	Linear representation of $\varphi$ for all combinations of prescription/recommended dose ranks given that the outcome was 0 MII retrieved. . . . .	42
3.7	Graphical representation of three individualized linear dose-response functions with three different slopes. The marked points indicated the dose needed to achieve 12.5 mature oocytes for each function. . . . .	53
4.1	Principal components of IDoser. . . . .	59
4.2	Graphical representation of loss evaluation for cases where $y < y^*$ (left) and where $y > y^*$ (right) . . . . .	61

4.3	Graphical representation of loss evaluation for cases where $y < y^*$ (left) and where $y > y^*$ (right) with additional rules considering maximum change allowed and a minimum change threshold. . . . .	62
4.4	$L$ across $d_{max}$ for La Marca, oracle and the proposed IDoser when used to dose in the validation dataset. The dashed line marks $L$ for the clinical practice dosing method. . . . .	67
4.5	Distribution of cases that need an increase of dose (red) or decrease (blue) for the validation dataset if $d_{max} = 450$ is allowed. . . . .	68
4.6	Distribution of doses for La Marca (blue), Clinical Practice (red) and IDoser (green) for the validation dataset with $d_{max} = 450$ . . . . .	69
4.7	Doses changes for the IDoser model with $d_{max} = 300$ . . . . .	70
4.8	Doses changes for the La Marca model with $d_{max} = 300$ . . . . .	70
4.9	Doses changes for the IDoser model with $d_{max} = 450$ . . . . .	70
4.10	Doses changes for the La Marca model with $d_{max} = 450$ . . . . .	70
5.1	Trial step-by-step flowchart. . . . .	75
5.2	SPIRIT flowchart of enrolments and assessments: Detailed timing for relevant events for participants during the randomized trial . . . . .	79
6.1	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. . . . .	91
6.2	Distribution of the differences in predicted probabilities for pregnancy and multiple pregnancy using Gradient Boosting Classifiers. . . . .	91
6.3	Logistic Regression and GBC pregnancy predicted probabilities plotted against maternal age and colored by embryo stage (blastocyst yes or no). . . . .	92
7.1	Distribution of cases by the number of gestational sacs observed around the 7th week of pregnancy in the whole database. . . . .	106
7.2	Distribution of cases by the number of gestational sacs observed around the 7th week of pregnancy after balancing strategies in the training database. . . . .	106
7.3	Distribution of changes on the number of embryos to be transferred recommended by IDoserET per number of gestational sacs observed. . . . .	108

A.1	Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 0 MII retrieved. . . . .	112
A.2	Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 6 MII retrieved. . . . .	113
A.3	Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 10 MII retrieved. . . . .	113
A.4	Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 15 MII retrieved. . . . .	114
A.5	Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 25 MII retrieved. . . . .	114

# List of Tables

2.1	Main contributions found in the current literature for AI driven solutions in different ART tasks. . . . .	24
3.1	Patient characteristics in the two databases used in the study. Values are expressed as average and SD. Variables were compared using Mann-Whitney U test. For proportions a 2-sample z-test was conducted. A p-value of $<0.05$ was decided as significant. . . . .	41
3.2	$\varphi$ values for every prescribed/recommended dose rank given that the result was 0 MII . . . . .	43
3.3	$\varphi$ values for every prescribed/recommended dose rank given that the result was 6 MII. . . . .	43
3.4	Mean absolute score ( $\Phi$ ) values plus 95% confidence interval (CI) for clinical dose rank prescriptions and model recommendations during development and validation phases. Statistical differences tested using the Wilcoxon signed-rank test. A p-value under 0.05 was considered significant. . . . .	45
3.5	Mean absolute $\varphi$ values ( $\Phi$ ) plus 95% confidence interval (CI) for clinical dose rank prescriptions and model recommendations during development and validation phases. Statistical differences tested using the Wilcoxon signed-rank test. A p-value under 0.05 was considered significant. . . . .	50
4.1	Summary statistics of development and validation databases. . . . .	63
4.2	Results of Iman Davenport's correction of Friedman's rank sum test of all methods tested across the 4 selected values for $d_{max}$ . . . . .	67
4.3	Ordered results from worst (left) to best (right) method in one vs one comparison across all $d_{max}$ values. Results extracted from post-hoc test with p-values adjusted by Finner's methodology. . . . .	67

6.1	Proportions of pregnancy and multiple pregnancy instances by group in patients who got transferred embryos at the cleavage stage . . . . .	89
6.2	Proportions of pregnancy and multiple pregnancy instances by group in patients who got transferred embryos at the blastocyst stage . . . . .	89
6.3	Results of the divided by type of model (Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier) and outcome (Pregnancy and Multiple Pregnancy). Mean effect shows the mean differences between chances predicted with DET minus chances predicted with SET. . . . .	90
7.1	Summary statistics of the embryo transfer database. . . . .	105
7.2	$L$ values for both the number of embryos prescribed by clinical practice and recommended by IDoserET. Differences stastically compared with the Signed-rank Wilcoxon test for $l$ values for each test case. A p-value under 0.05 was considered significant. . . . .	107
7.3	Number of embryos recommended by IDoserET as compared to clinical practice (rows) separated by the real outcome observed (columns). The cells where the changes are considered good are colored in green, in red the changes considered bad, and in yellow those where results are considered uncertain. . . . .	108
A.1	Score values for every prescribed/recommended dose rank given that the result was 0 MII . . . . .	111
A.2	Score values for every prescribed/recommended dose rank given that the result was 6 MII. . . . .	111
A.3	Score values for every prescribed/recommended dose rank given that the result was 10 MII. . . . .	112
A.4	Score values for every prescribed/recommended dose rank given that the result was 15 MII. . . . .	112
A.5	Score values for every prescribed/recommended dose rank given that the result was 25 MII. . . . .	112
B.1	Posthoc test of differences in individual losses between explored methods and clinical baseline capping $d_{max}$ at 300, p-values adjusted using Finner method and marked with * when under 0.05. . . . .	117

B.2	Posthoc test of differences in individual losses between explored methods and clinical baseline capping $d_{max}$ at 350, p-values adjusted using Finner method and marked with * when under 0.05. . . . .	117
B.3	Posthoc test of differences in individual losses between explored methods and clinical baseline capping $d_{max}$ at 400, p-values adjusted using Finner method and marked with * when under 0.05. . . . .	118
B.4	Posthoc test of differences in individual losses between explored methods and clinical baseline capping $d_{max}$ at 450, p-values adjusted using Finner method and marked with * when under 0.05. . . . .	118

# Acknowledgements

First and foremost I want to express my gratitude to my supervisors. To dr. Rita Vassena, for giving me the opportunity of stepping out of the embryology laboratory to dedicate myself to the intersection of AI and ART. Without your trust in me, and your drive for innovation, this thesis wouldn't exist. To dr. Josep Lluís Arcos, for your continued support and (very) patient teaching. Your ability to look at things from different angles has been key to reach the solutions we present here. To dr. Jesús Cerquides Bueno, also for being so patient with me, even when my mathematical notation was messy or even absent. Your (always on point) criticism has made this thesis something that I can be proud of. I have learnt so much from both of you that, maybe some day, this embryologist will believe herself a real data scientist. To all three of you, I cannot overstate how thankful I am, for the opportunity granted, the unwavering support, and for how easy you have made this process for me.

To Pau Olivés Tarrés, for helping me with the last leg of this thesis, even when having to deal with my messy notebooks.

To the IIIA-CSIC, for granting me the opportunity to meet so many interesting researchers from very different disciplines, and to learn from them, in a very easy-going environment.

To the Universitat Autònoma de Barcelona and to the industrial doctorate grant program from the Generalitat de Catalunya. Without the grant I would not have been able to return to my alma mater, in a personal moment where I really needed to return to my roots.

To Eugin Group and CIRH, for giving me the opportunity to form myself as a clinical embryologist, and then leaving me room to grow to contribute to the team in a whole different way. To dr. Maria Jesús López Martín and dr. Daniel Mataró Marsal, for your deep knowledge on FSH dosing. To the I+D team, specially dr. Mina Popovic, for being always there for me, cheering me on and letting me shower you with questions whenever I was insecure.

And to my family and friends. To my parents, who always pushed me to be better and go farther. Without your efforts and deep love I wouldn't be here to begin with. To my friends, my found family. You give my life light and laughter, and have been steadfast always by my side. I am the person I am today thanks to you all. But very specially, to my wife, for always believing in me,



much more than I would. For putting up with me when I was tired, sad, anxious, insecure. For giving me insight on communicating complex concepts properly, and hearing me rambling on my thesis. You and our son are the reasons that lead me to want to be better. Thank you.

# Chapter 1

## Introduction

### 1.1 Thesis motivation

Every clinical embryologist works day-to-day striving to achieve success for every patient. As success in this context means a healthy baby at home, it is easy to understand why professionals in the field of human assisted reproduction are heavily invested in accomplishing it for each and every patient in their care. 45 years after the birth of Lousie Brown (the first baby born via in vitro fertilization) great technical advances have been achieved, however success rates per IVF cycle still remain around 30%. This directly translates into many patients not reaching their desire of adding a new member to their family at their first attempt, second, or sometimes ever. From the professional point-of-view, often these poor outcomes can give way to frustration, as the hard gained expertise and diligent work on the case has not yielded positive results. Sometimes it is partly expected, due to the characteristics of the patient, such as biomarkers already indicating a poor prognosis, but there is a non-negligible percentage of failures where the prognostic was good. Always, and especially in these cases, professionals feel the pressure of understanding the cause of the unsuccessful treatment in order to improve the patient chances in the next attempt.

The field is continuously endeavoring to advance basic and clinical knowledge that can lead to technical advances. In parallel, artificial intelligence (AI) has also been progressing at an enormous pace, with methods like machine learning (ML) and its subclass deep learning (DL) garnering widespread interest and recognition from both the general public and the scientific community. Ever since the contentious victory of Deep Blue, an AI-powered chess computer, over the reigning world champion Garry Kasparov in 1997, the notion of AI surpassing human intelligence has been in the collective mind. Even though AI singularity (the event where AI could surpass human intelligence in all aspects) is perhaps a future event, certainly AI, and especially ML, can detect previously undetected patterns in high dimensional datasets. This possibility caught the attention of many researchers dealing with large sets of complex data, where detailed analysis

needs high computational power. This applies too to healthcare disciplines like assisted reproductive technologies (ART). Most professionals of this field would ponder whether AI could help us understand and/or improve those cases where success is not achieved.

In ART, as in many healthcare specialties, information on the relevant biomarkers of the patient (i.e. age, previous medical history, pertinent blood tests, results of treatment, etc.) are gathered and stored safely in the points of care. This information is commonly used to run retrospective studies, but there are many questions whose answers need to, ideally, check or adjust for many related variables. Here is where AI, and especially ML, shines, as it is capable of processing many variables and cases and detect previously undetected correlations.

An experiment on using ML to predict the result of a controlled ovarian hyperstimulation (COH) was performed before this thesis, as COH is a key step in an in vitro fertilization (IVF) process, where a failure can compromise the result of the cycle. The model, trained with patients characteristics like age and body mass index (BMI), ovarian markers, and the first dose of follicle stimulating hormone (FSH), was intended to predict the number of mature oocytes retrieved after ovum pick-up (OPU). Prediction quality scores during the test phase of this first experiment were good, and an abstract of the study was presented for ESHRE 2018, where it was accepted as an oral presentation (“Abstracts of the 34rd Annual Meeting of the European Society of Human Reproduction and Embryology” 2017).

The idea behind the study was not to predict mature oocytes exactly, but rather to use it as a COH protocol simulator, for clinicians to be able to adjust dose in-silico depending on the outcome predicted for the individual patient. But when work on the project was resumed and the first in-silico trials were performed, the trained model predicted consistently less oocytes when FSH dose was increased for each individual patient. Which was, of course, far from reality. This could be easily visualized with plots like SHapley Additive exPlanations (SHAP) dependence plots, like the example in Figure 1.1. This kind of plot provides a visual and intuitive representation of the dependencies between variables, helping to explain the model’s predictions and identify influential features.

In the case of the model here explored, the variable “stim”, which identified the starting dose of stimulation drug (FSH), was marked as blue. This exploration was repeated with multiple different fictional examples, finding that the FSH dose variable was always coded in blue. This meant that the model understood the dose-response relationship as a negative one.

After close examination, it became clear that the database distribution was the culprit. That, and that AI is really not that intelligent, and only learns from the data we provide it. Thus, if in the dataset patients with lower doses always obtain more oocytes and those with higher doses less, that is what the model learns. Of course, patients that receive high doses are those with poor prognostic biomarkers, and on the other side of the spectrum the contrary happens. Clinicians are adjusting

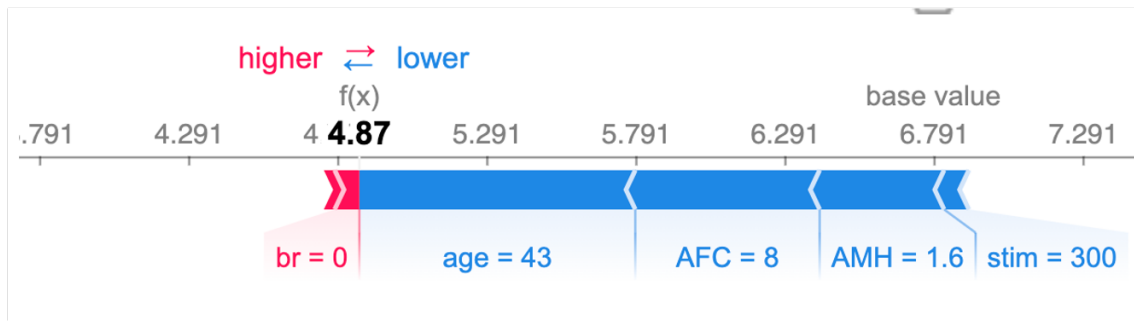


Figure 1.1: SHAP dependence plot for a fictional example patient. In red factors that increase the value of the number of oocytes predicted. In blue factors that decrease the prediction. Other variables present in this plot are related to ovarian reserve or patient age at the moment of treatment.

the treatment taking the individual characteristics of the patient in order to get an optimal result and minimizing risks. That is human intelligence, using expertise garnered by past experiences and deep knowledge of the field to decode what an AI model alone cannot.

Confounding, which is the phenomenon present in the dataset that made the first model to be clinically imprecise, happens when a variable affects both treatment allocation and the outcome under study. In this case, as confounding was not dealt with, the model was understanding that the lower the dose, the higher the outcome. Experts on the process can easily say that if adjusted by biomarkers, actually, that trend reverses, and the relationship should be positive up to a point of saturation (more on that on the next chapter). A prospective randomized and controlled study (RCT) would deal an uncounfounded dataset. But RCTs are expensive and time-consuming, so they are rarely performed. Observational or retrospective datasets are the common source of data in clinical setups available for research and/or AI projects. This kind of dataset tends to be confounded, limited in treatment variation (as clinical protocols are followed closely), and usually suffer from variable annotation quality and/or completeness. There are ways to overcome confounding, which will be covered in the next chapter, but they often struggle when data is not complete and varied enough.

A human expert is able to look at a confounded dataset, that does not contain enough information for the AI model to understand the real dose-response relationship, and use their knowledge to fill-in the gaps. The opportunity then, is to combine the high computing power of AI methods and the clinical and field knowledge not present in observational datasets. Then, clinically robust dosing models should be obtained, together with the ability to analyze if their decisions are set in the right direction, and thus may provide an actual benefit for the patients.

## 1.2 Research questions

COH is, as stated before, the first step in many IVF treatments, and its objective is retrieving an optimal number of mature oocytes. This usually means a high enough number as to maximize

the possibilities of a pregnancy, but not as high as to endanger the health of the patient. The first dose of FSH in a COH protocol will recruit a given number of follicles; those will be surgically punctured at the end of the stimulation in order to retrieve the oocytes. As once initial recruitment is achieved the number of follicles hardly changes, it is clear that the selection of that first FSH dose is key not only for COH, but for the whole cycle's success.

Unbiased outcome data of a dose-response relationship is ideally obtained from randomized and controlled prospective studies. This type of studies are, in fact, economically expensive and time-intensive, which leads to researches to rarely being able to invest in them. Additionally, they are only performed with a justified cause, like testing a new treatment compared to standard practice. In other words, performing an RCT just to obtain a varied database would mean treating randomly patients without a real clinical cause, which is simply bad practice. As such, observational data stored from past treatments is often the realistic source of information for dose-response analysis for drugs already in use in clinical settings.

This reasoning drives the question that this project tries to answer first, which is:

- **Q1:** *Is it possible to improve clinical FSH dosing policy for Controlled Ovarian Hyperstimulation using only historical data?*

Given the results of the project presented in ESHRE 2018, and after analyzing and learning from them, it is clear that to really answer this question we first need to state what an improvement really means. Good values in typical prediction scores as correlation between predicted and real variables do not directly imply a clinical improvement for the patients. Thus, stopping performance analysis there will not ensure that the dosing model is clinically robust. Then, in order to be able to answer the first question, we need to answer:

- **Q1a:** *How can we analyze a dosing model's performance before clinical intervention?*

Designing a trustworthy system of evaluating the performance of an AI driven dosing model implies the codification of professional field knowledge absent in historical data. Such a system will enable realistic analysis of dosing models. This is specially relevant, as nowadays for any dosing model to be actually used in a clinical setting it needs to be tested in a randomized controlled trial (RCT). Selecting the best model according to a clinically driven evaluation can potentially facilitate the approval of such a trial, and safeguard the health of patients during the trial.

Once this is possible, the next question is:

- **Q2:** *Can we extend this methodology to other dosing problems?*

With the first two questions answered, lessons learned can be applied to other similar problems in the IVF process. One such a possibility is the selection of number of embryos for

transfer. Selecting the right number given the characteristics of the patient, her cycle and the embryos should lead to the desired outcome: a single clinical pregnancy. The objective would be to develop a dosing model able to improve the rate of success of this event, while minimizing risks of no pregnancy and multiple pregnancy.

Questions 1 and 1a will be answered in Chapters 3 and 4, and a method to answer question 2 is presented in Chapter 4. Its application is reviewed in Chapters 6 and 7.

### **1.3 Ethical aspects of this project**

Obtaining ethical committee acceptance for study protocols is essential to ensure the protection of participants' rights, maintain scientific integrity, and uphold ethical standards in research. As such, EU and Spain regulatory bodies govern over the necessary permits for clinical and observational studies to be performed in their territories. In Spain, under the regulation Real Decreto 957/2020, from the 3rd of November 2020, observational studies must obtain the approval of an ethical committee beforehand.

All datasets used in this thesis are observational and have been retrieved and analyzed after approval from an Ethical Committee.

For the FSH dose-response project, permission to conduct the study was obtained from the Ethical Committee for Research of Eugin on 20 October 2020 (approval code: ALGO2).

As for the selection number of embryo for transfer project, permission to conduct the study was obtained from the Ethical Committee for Research of Eugin on 24 March 2021 (approval code: SINGLE).

Additionally, the RCT protocol presented in Chapter 5 has also been reviewed and given approval from the Ethical Committee for Research of Eugin on 8 March 2023 (approval code: ALGO3).

### **1.4 PhD structure**

This thesis is structured as follows:

- Chapter 2 covers the current state of the art. Starting with general concepts of AI in health-care, the many ways it can be implemented and examples of preliminary success, its main pitfalls and some further details on the focus of this thesis, dose-response modelling. It will follow-up with details of AI in ART, covering first aspects of natural conception and infertility, and how ART and specifically, IVF tries to improve chances of pregnancy. Lastly,

examples of AI implementations on multiple steps of the IVF process will be reviewed, with especial focus on COH and the selection of number of embryos for transfer, together with a commentary on RCTs performed up to date.

- Chapter 3 includes details on methodology and results of two first iterations on FSH dosing models for COH, together with an ad hoc score function designed to evaluate how good the doses recommended by the models are in comparison with the actual clinical prescriptions.
- Chapter 4 recounts the third and final iteration of the FSH dosing model, alongside the description of the generalizable methodology for training and evaluating models that incorporate field knowledge, which was named IDoser.
- Chapter 5 describes the complete design of an RCT protocol for the final IDoser model for FSH, which will test the hypothesis of non-inferiority compared to the current clinical protocol related to the outcome of number of mature oocytes retrieved.
- Chapter 6 introduces a first approach to a second point of interest of this thesis related to IVF, the selection of the number of embryos for transfer. Out-of-the-box ML models are intentionally implemented naively to later review whether their behavior is clinically robust.
- Chapter 7 recounts general conclusions of the thesis, reports the contributions and publications originated by it, and outlines future lines of work of this line of research.

Next, the concepts of AI, ART, and the literature where both meet will be reviewed.

## Chapter 2

# State of the art

### 2.1 AI in Healthcare

The term artificial intelligence, defined as the science and engineering of making intelligent machines, was coined by John McCarthy, Marvin Minsky and Claude Shannon in the Dartmouth Conference in 1956; made possible by the work and advancements of many great scientists such as Alan Turing and Ada Lovelace. Since then it has continued evolving and gathering attention, as there are many knowledge fields that stand to benefit from its diverse implementations. One of such fields is healthcare where, since the introduction of electronic health records (EHR), relevant information has been accumulating, generating extensive databases that could be exploited.

The remarkable potential of AI in healthcare has already been validated through promising applications in several aspects, which include:

- **Medical diagnosis:** EHR databases represent a massive source of information regarding medical history of individual patients, which includes results from blood tests, imaging, treatment plans often with their outcomes, etc. All this data has been historically used by clinicians to diagnose patients, and AI has been applied in the same direction in the hopes of improving accuracy, swiftness and robustness of such diagnosis. Some examples are its applications in cancer (Melarkode et al., 2023; Kakotkin et al., 2023; Stan-Ilie et al., 2023; Pesapane et al., 2023) and heart disease (Kampaktsis et al., 2023; Denysyuk et al., 2023).
- **Remote monitoring and telemedicine:** Ever since the wearable devices have been widely accessible and priced reasonably, remote monitoring of patients basic vitals like heart rate or oximetry has been a realistic option and thoroughly explored. Together with invasive or non-invasive medical devices specially devised to monitor particular illnesses like blood sugar monitoring for diabetes, or connected ventilators for chronic obstructive pulmonary



disease, a wealth of information has been gathered. This has led to further AI models trained in order to accurately register relevant events or to obtain thresholds that determine need of intervention/hospitalization (Pépin et al., 2022; Gautam et al., 2022; Arfan Ahmed et al., 2020). This advancements have been specially relevant during the COVID-19 pandemic, where face-to-face interactions had to be drastically minimized, and healthcare experienced a paradigm shift to ensure safety of both patients and clinicians. Remote monitoring and telemedicine (aided in some cases with AI chatbots [Aggarwal et al., 2022]) has proven extremely useful both during the pandemic related restrictions, and after (Adnan Ahmed, Charate, and Pothineni, 2020; Nazir et al., 2022; Motwani, Kumar, and Pawar, 2020).

- **Personalized treatment:** The wealth of data on illness related biomarkers can also be used to improve individualized treatment. During the last years a significant rise in research output in this regard has been observed for multiple diseases (Eskofier and Klucken, 2023), specially in cancer treatment (Shao et al., 2023; Gui et al., 2023) where genomic and transcriptomic data is being included in predictive models with good results. Other examples are hypertension (Visco et al., 2022) or sleep apnea (Brennan and Kirby, 2023) management. Another compelling way of personalizing treatment is the incorporation in AI modeling of patients' own views of their symptoms, or patient-reported outcome measures (PROMs). The inclusion of this kind of variables humanizes AI models, and guides training with a more comprehensive picture of the patient's condition. Essentially, giving patients and clinicians the opportunity to make the best healthcare decisions together (Cruz Rivera et al., 2023).
- **Drug discovery:** AI is emerging as an increasingly important tool to improve the efficiency and effectiveness of the drug development process. Its applications are varied, from biological target identification, prediction of protein-protein or drug-protein interactions, drug repurposing, to virtual screening of molecules in expansive docking libraries (Potlitz, Link, and Schulig, 2023; Das and J, 2023; Raza et al., 2022; Yoo et al., 2023; Sarkar et al., 2023).
- **Healthcare operations:** Further uses in healthcare include managing staff schedules and resources, reducing wait times, and improving patient flow (N. Rozario and D. Rozario, 2020; S. Jiang, Liu, and Ding, 2023; Hosseini et al., 2023).

Although all of these uses of AI for improving healthcare outcomes and processes are certainly promising, there are rising concerns on how they are developed and or implemented. AI and, specially machine learning (ML) and deep learning (DL) are highly appreciated due to their capability to process high-throughput databases, and in the process detect unknown patterns that can help predict better certain outcomes. DL, specifically, is extremely useful to process imaging data, as it is capable of automatically pre-process it in order to extract relevant information. But all these advantages come with some cons.

### **Interpretability and explainability of AI models:**

Transparency of several ML algorithms and specially DL ones can range from high to very low. Some of them are called black boxes because of the high complexity architecture of the resulting trained models, making it difficult for users and AI experts to understand how they make a specific decision. Although the performance of the model can be highly accurate, users can find it difficult to trust them due to their opacity. Here, two popular concepts in the AI field are relevant:

- **Interpretability:** In the context of AI, this concept refers to the degree to which the internal mechanisms and decision-making processes of a predictive model can be comprehended. It encompasses acquiring insights into how the system reaches its conclusions or predictions, such as discerning the influential input features or comprehending the connections between the data and the output. The objective of interpretability is to establish a level of transparency and understandability in the functioning of the AI model. A good example of a highly interpretable model is a linear regression, where output is linked to input variables through a set of coefficients in a linear function. Any variable linked to a negative coefficient can be understood as a negative influence for the outcome, and the contrary for any positive coefficient linked variable.
- **Explainability:** This concept defines the effort to take an ML model and explain the behavior in human terms. With complex models (for example, black boxes), it is certainly challenging to fully understand how and why the inner mechanics impact the prediction. However, through model agnostic methods (for example, partial dependence plots, SHapley Additive exPlanations (SHAP) dependence plots, or surrogate models) it is possible to discover meaning between input data variables and model outputs, which enables the explanation of the nature and behavior of the AI/ML model.

In summary, an interpretable model can also be very easily explained, but not all explainable models are fully interpretable. Some black box models can be explained via model agnostic methods in an effort to increase their transparency, but achieving full understanding of their inner work remains challenging.

This relates with the following disadvantage: the dangers of biased databases. ML models learn from the data they are fed on, hence if the data is biased, the model will capture that trend too and reproduce it in its predictions. This can be specially detrimental if the trained model is not transparent, as it will be challenging to detect the inherent bias rapidly.

This is exemplified in the study by Caruana et al., 2015 describing setbacks found in previous works (Cooper, Aliferis, et al., 1997; Cooper, Abraham, et al., 2005). Particularly the bias found in models trained to predict probability of death for patients with pneumonia. Two models were

tested: a neural network (black box) and a logistic regression. Even-though the neural network performed much better than the logistic regression, it was considered too dangerous to be applied to a real clinical setting, as the interpretation of the regression raised warning flags on bias that most probably would also have been introduced in the neural network. Specifically, the logistic regression model predicted that patients with asthma had higher probability of survival, which is counter intuitive. Of course, the data did show this trend, due to a higher intensity of treatment that these patients experience, on average, because of their health condition, ultimately leading them to higher survivals in comparison with non-asthmatic population. From the start, the clinical decision to aggressively treat them is explained by their high-risk condition as asthmatic patients. Any expert in the field can pinpoint that reasoning, but ML models lack this context if not codified in the data shown to them during training, and so introduce bias in their architecture. This example is relevant for all ML applications, but particularly to those in healthcare, as biased decisions can have pernicious consequences on the health of patients. In this instance an interpretable algorithm was critical to the detection of bias before harm could be done.

Unfortunately, bias has not always been detected before deployment of algorithms. Obermeyer et al., 2019, describes an algorithm commercially employed to predict risk scores for patients which was heavily racially biased against black populations. At a given risk score, a black patient was considerably sicker than a white patient. This, of course, led to less investment in their needed treatments, to the detriment of their health. Adequate countermeasures were put in place after detection, but significant harm had already been done.

Research is being conducted on the revision of interpretability of published models (Xu et al., 2023), and to provide tools to easily interpret and/or explain trained models (Ribeiro, Singh, and Guestrin, 2016).

#### **Data availability:**

Another challenge that arises when implementing AI in healthcare is data availability. For ML and DL models to be able to learn properly huge databases are desirable. But as clinical data is considered, and rightly so, confidential, it is highly protected. As such, data tends to be isolated in individual centers, and only big hospitals, groups or public healthcare networks are privy to sufficiently big databases. Additionally, patient consent is key for data privacy, and there has been intense debate on this regard, as NHS came under heavy criticism after handing over patients' data without explicit consent in order to develop an app in collaboration with DeepMind, an AI subdivision of Google (Powles and Hodson, 2017). With newer stricter legislation, these exchanges have been made more difficult, protecting patient privacy, but slowing down possible cooperations.

#### **Clinical validation of AI models:**

Finally, the literature on clinical validation of AI models is sparse at best, as Randomized Controlled Trials (RCT), the gold standard in hypothesis testing for medical interventions, are time and

money intensive. Many projects remain at the pre-clinical stage for that reason. These drawbacks are thoroughly covered in Khan et al., 2023.

Trust is paramount for any AI model to be adopted by professionals, and this trust is earned by addressing the challenges mentioned above with the care they deserve. Ensuring that models are as little biased as possible, thus consistent with field knowledge, and transparent enough for users to understand how they make their decisions will increase user confidence (Obermeyer et al., 2019; Khan et al., 2023). This has also to be supported by an easy-to-use platform to query the model, implemented in a way that does not disrupt the clinical workflow.

### **2.1.1 Dose-response modelling**

In this thesis we focus on personalized treatment, specifically in optimal drug dose selection. In order to select an optimal drug dose, a response or outcome needs to be defined as the endpoint to reach via the selected dose of drug. The relationship between dose and response, also known as exposure-response relationship, refers to how much an organism responds to a stimulus or stressor (often a chemical or drug) based on the amount of exposure or dose it receives after a specific period of time. Dose-response curves are used to describe these relationships (Hayes, Wang, and Dixon, 2020). Understanding the relationship is critical for any dosing protocol or policy to be effective. Modelling it via curve functions is how historically this problem has been approached (Calabrese, 2016).

The dose-response concept is fundamental in toxicology and pharmacology, and its modelling via curves is used extensively in drug development. The shape of a dose-response curve can provide insights into the underlying mechanisms of the substance's effect, and be used to predict the effectiveness and toxicity of drugs or stimuli.

These curves can take multiple shapes. One of the main theories underlying dose-response curves is the idea of hormesis. Hormesis is the phenomenon by which exposure to a substance at high doses is toxic, but at lower doses it has a beneficial effect. Plotting the substance's benefit can result in a U-shaped curve, where very low doses of the drug don't have an effect, medium- to-low doses have a positive effect, but higher doses become increasingly harmful (Calabrese and Baldwin, 2002).

Another important theory in dose-response modelling is the threshold model. It suggests that there is a certain threshold of dose below which there is no clinically significant or detectable effect, and above which the response increases in proportion to the dose. This can take the shape of a linear dose-response function, where the effect increases in a straight line as the dose increases. Related to this model, is the linear non-threshold model (LNT), used habitually in the radiation science (Sacks, Meyerson, and Siegel, 2016). Here, effects are always deemed harmful, even if dose is low.

Finally, the concept of saturation can also play a role in dose-response curves. Saturation occurs whenever a drug or substance reaches a maximum response at a certain dose, beyond which further increases in the dose have no additional effect. This can be represented by a sigmoidal dose-response curve, where the effect initially increases rapidly as the dose increases, but then levels off as the substance reaches its maximum effect. Many biological response curves can be closely approximated as a sigmoidal shape, as saturation plays part of those processes, via, for example, occupation of all available specific receptors for the substance. The Hill equation, non-linear logistic function composed by 4 parameters, is frequently used to fit these relationships (Gadagkar and Call, 2015).

Model-informed precision dosing (MIPD) aims to enhance the outcomes of drug treatment for patients by attaining the ideal equilibrium between efficacy and toxicity that is tailored to the individual patient. MIPD includes both pharmacometrics and AI-driven approaches (Keizer et al., 2018; Darwich et al., 2017).

On the one hand, quantitative dose-response analyses are often referred to in general as pharmacometrics. Pharmacometrics (PX) can be described as the science that develops and applies mathematical and statistical models to analyze, understand, and predict a drug's pharmacokinetic (PK), pharmacodynamic (PD), and outcome behavior (Barrett et al., 2008).

On the other hand, whenever PX methods are not applicable, ML methods have been proposed. Specifically, approaches that rely on the concept of causal inference. Causal analysis aims to infer the causal effect of a specific treatment or action under certain conditions for a particular outcome (Pearl, 2010). Its capability to model the causal relationship between treatment and outcome, and to condition it on a set of covariates, is of great relevance for approximating individualized dose-response functions.

In the next subsections we will review briefly both approaches. In Chapter 4 a novel methodology is presented, designed to be used whenever PX and causal inference methods are not applicable.

## **Pharmacometrics**

PX applies mathematical and statistical models to understand a drug's PK and PD. PK focuses on the relationship between the dose administered and its concentration in different body compartments at specific time points. Conversely, PD focuses on what the drug does to the body, namely in the exposure to effect relationship (Lewis and Wakefield, 1999). PK/PD studies where the main bio-markers are included constitute the most physiological approach to MIPD. In its ideal form, it will result in obtaining well fitted models for both PK and PD of the studied drug. Drugs approved for clinical use often have a published PK/PD model, derived from phase III clinical trials.

Pharmacokinetic/pharmacodynamic (PK/PD) studies, used to fit the models to prospectively col-

lected data, constitute an important step in drug development and clinical approval. These studies inform the design and execution of Randomized Controlled Trials (RCTs), in which the drug is tested in clinical conditions. Clinical dosing protocols are then derived from the results of RCTs to guide practice, ensuring it is safe and optimizes the therapeutic effect. Ultimately, results range from acceptable to almost optimal, depending on the fit of the PK/PD models on the final objective population and real clinical outcome. The final unexplained variability constitutes the aim of further research. This variability is mainly related to the target population distribution and its accompanying comorbidities.

More often than it is desirable, unfortunately, these models are not applicable to all patients. First, phase III PK/PD studies tend to exclude most or even all comorbidities known to affect the drug, although it is relatively common for individuals to have comorbidities (Gonzalez et al., 2017). Second, only a relatively selected patient population tends to be included in these trials and their precursors (phase I and II), usually healthy male adults. Third, the application of these models assumes that the target population parameters have the same distributions as the study sample, often an incorrect assumption due to, for instance, socioeconomic status, genetic and ethnic dispersion and geographical distribution, amongst others (Keizer et al., 2018). Finally, it is not uncommon to find PK/PD models fitted for outcomes that are different from the objective of clinical interest, or unrelated to key bio-markers known to affect the individual dose-response function variability.

While, generally, this system allows for most patients to receive an adequate dose of drug, dose protocols or policies may not be fitting all patients evenly. Whenever this is the case, clinicians use their own experience and the published literature to fill in the knowledge gaps, however this is still not optimal due to quality issues in published research, patient group coverage, and unreported bias. There are, however, ways to adapt these initial models to real-world clinical use. Some examples are the use of non-linear mixed methods (Lewis B. Sheiner and Ludden, 1992; L B Sheiner and Steimer, 2000), physiologically based PK models (PBPk) (Jones et al., 2015), Bayesian methods (Lewis B. Sheiner and Beal, 1982; Darwich et al., 2017; Hamberg et al., 2015), and traditional PK/PD methods combined with machine learning (ML) (McComb and Ramanathan, 2020). However, for the dosing model to be improved, all of these approaches require that a covariate/bio-marker linked PK/PD model to be improved is available as a starting point. Alternatively, a new PK/PD model connected to said covariates can be developed, yet this approach is both computationally and labor-intensive, while often requiring data on drug concentrations in the body after treatment, which may not be available (McComb, Bies, and Ramanathan, 2022; Koch et al., 2020). Regardless, this approach implies either using new prospective data, or very diverse and complete observational datasets. The first option involves collecting new data, which is expensive in both time and resources. The second option remains difficult to define. While clinicians do adapt protocols when necessary, they try to not deviate from established safety and effectivity standards, generating as a result data with sparse diversity for dose allocation. Additionally, as already men-

tioned, there is data that is potentially not registered in clinical practice, like blood concentration of the drug after its administration.

Nevertheless, the reliance of these methods on physiological and pharmacological concepts is a great advantage for the explainability and trustworthiness of the resulting models.

## Causal inference

The goal of MIPD is to optimize the effectiveness and minimize the toxicity of dose selection for individual patients by achieving a balance tailored to their specific needs and characteristics. In the previous subsection we have presented how PX methods could reach a solution for MIPD, and their limitations. In this subsection we will review causal inference and its potential advantages and disadvantages.

In order to understand causal inference, first, we need to define causation: "A variable  $X$  is a cause of a variable  $Y$  if  $Y$  in any way relies on  $X$  for its value" (Pearl, Glymour, and Jewell, 2016). Thus, causal inference is the science that tries to determine causation between two variables (treatment and outcome), models it and uses it to estimate outcome in counterfactual scenarios where the treatment variable is different than the one present in reality.

As dose-response relationships indeed reflect a causal connection between treatment and outcome, causal inference, at first glance, should indeed be a good fit for dose-response estimations. In contrast, machine learning methods work mainly by seeking correlation between input variables and output variables. But presence of correlation does not always mean that a causal relationship is also present (Scholkopf et al., 2021). Hence, a method closer to the real relationship exposure-outcome, and more interpretable is, indeed, causal inference.

Causal analysis also deals with the concept of confounding. Confounding can be defined as the presence of spurious association between two variables (i.e. treatment and outcome) due to the influence of external factors (Pearl, 2000). A clear example of this effect is the Simpson's Paradox (Simpson, 1951; Blyth, 1972), that describes the phenomenon where a variable ( $X$ ) increases the probability of another ( $Y$ ) in a population ( $p$ ), but simultaneously decreases the same probability of  $Y$  for every subpopulation of  $p$  (Pearl, Glymour, and Jewell, 2016). This can be easily understood with the example shown in Figures 2.1 and 2.2. There, the results of a study where weekly exercise and cholesterol levels are measured for multiple age groups are shown, first unsegregated and then segregated by age group. Looking at Figure 2.1 one may conclude that an increase in weekly exercise is associated to high cholesterol, maybe even causing it. But common knowledge would indicate that this is a mistake. In fact, when the same data points are segregated by age group (Figure 2.2) the real causal relationship is revealed: exercise lowers the levels of cholesterol. The dataset of this study could be labelled as confounded, and the culprit or confounder variable is age

in this case. A confounder can be defined as any variable that affect both treatment/intervention and effect/outcome (Pearl, 2000).

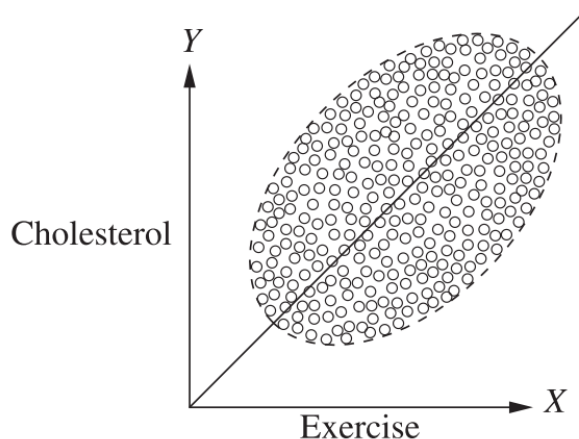


Figure 2.1: Results of the exercise–cholesterol study, unsegregated. Taken from Pearl, Glymour, and Jewell, 2016.

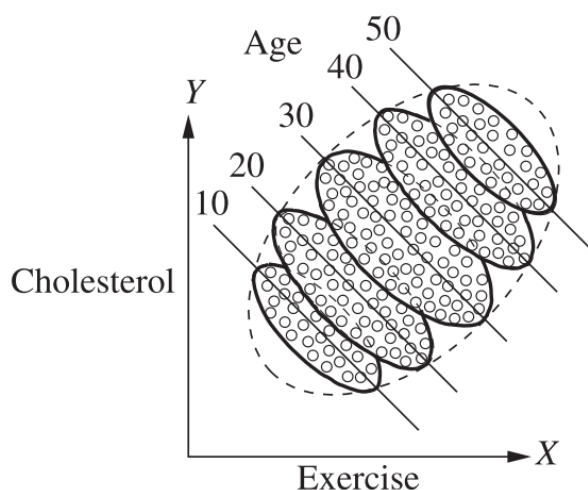


Figure 2.2: Results of the exercise–cholesterol study, segregated by age. Taken from Pearl, Glymour, and Jewell, 2016.

The confounding effect is very common in dose-response relationships in healthcare, given that for any specific problem it is usually possible to identify several biomarkers that affect both how the treatment is allocated and the final outcome that needs to be measured.

Causal inference formally defines assumptions of causal relationships in a dataset via structural causal models (SCM), which are related to a graphical causal model, generally represented as a directed acyclic graph (DAG). These graphs allow for high transparency on how a causal model associates a variable to another in the studied datasets, and allow for the identification of key confounders that need to be adjusted by in order to see the unconfounded relationship between



treatment and effect. A simple example of a DAG can be found in Figure 2.3.

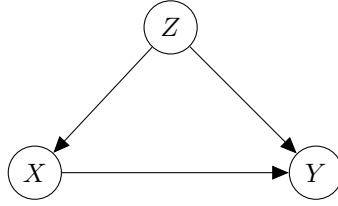


Figure 2.3: Directed Acyclic Graph showing a confounder ( $Z$ ) affecting both treatment ( $X$ ) and outcome( $Y$ ).

Recently, these two disciplines, causal inference and ML, that have been evolving separately, are being joined in an effort to exploit the strong points of both (Scholkopf et al., 2021). In the specific case of dose-response relationship modelling, ML and causal inference methodologies are being combined for single observational data, as presented by Bica and Jordon, 2020. For binary treatments the propensity score (probability of an individual of receiving certain treatment) is used to adjust for treatment selection bias.

For multiple or continuous treatments this concept is translated to the generalized propensity score (GPS) (G. Imbens, 2000; Hirano and G. W. Imbens, 2005). This score is used to weight samples while estimating the outcome value. Unfortunately, propensity score models must be very well determined and can be numerically unstable due to extreme propensity weights (Bica, Alaa, et al., 2021).

Recent methods to ameliorate the same problem include kernel functions to estimate the GPS (Colangelo and Lee, 2020; Kallus and A. Zhou, 2018) and Doubly Robust (DR) ML to estimate outcome values. Other works discretize the treatment space (Cai et al., 2020; Schwab et al., 2019), or use generative adversarial methods (Bica and Jordon, 2020). These are very good approaches for estimating dose-response relationships but require making two assumptions key for every causal inference analysis (Pearl, Glymour, and Jewell, 2016):

- Positivity or overlap: every individual has non-zero probability of receiving every treatment option.
- Unconfoundedness: all treatments and outcomes affecting variables are accounted for.

In clinical settings, these assumptions can be very challenging to fulfill. As stated before, clinicians seldom deviate from the clinically accepted dosing policy, thus they tend to dose similarly patient with comparable characteristics. This hinders the positivity assumption, as it is frequent to encounter dose-response problems where there are groups of patients with no data in certain dose ranges. Furthermore, it's not always possible to adjust for all confounding variables in a uniform manner across all cases. This is because clinicians have personal and practice preferences

for different biomarkers or may administer more or less comprehensive test batteries depending on factors such as their expertise, financial considerations, or patient requests. As such, the unconfoundedness assumption is also difficult to comply with.

### **2.1.2 Conclusions**

ML methods have been applied with varying success in several specialties, like cancer chemotherapy treatments (Yang et al., 2023). Caution must be used before the implementation of ML models in clinical practice. The Ethics Guidelines for Trustworthy AI, formulated by the EU Commission’s High-Level Expert Group on Artificial Intelligence in 2019, mandates that machine learning models should not only guarantee fairness and minimize bias but also ensure accountability and transparency (Commission, 2022).

The potential of the harmonious matching of PX and ML methodologies in order to achieve ideal MIPD is very significant, but will require close collaboration between clinicians, pharmacologists and experts in ML (Poweleit, Vinks, and Mizuno, 2023). There is a consensus on the need for further research, as there are specific aspects of its application to healthcare that need to be overcome (Ota and Yamashita, 2022). On one hand, healthcare deals with the well-being of people, and as such any AI implementation in clinical practice needs to be closely reviewed. On the other hand, clinical datasets are prone to be limited in sample size, but with many relevant features. Dose-response relationships known to be modulated by those many relevant features need big sample sizes in order to fit adequate models, as models fitted with small sample sizes can detect spurious correlations. PX and causal inference methods are very good candidates to estimate dose-response relationships, but need either data obtained after a prospectively randomized trial, or observational data with enough cases, all known confounders annotated, and variability in the allocation of dose for similar patients. In reality, observational clinical datasets are the main source of data to optimize dosing policies, and are far from the minimum requirements needed for either PX or causal inference methods to work properly. Hence, other methods are needed.

## 2.2 AI in ART

We will structure this section in two parts:

1. **General overview:** Where AI-driven methods tailored for ART are reviewed overall.
2. **Dose recommendation in IVF:** Where a zoom in in the areas in which this thesis is interested is performed.

First, we will review the general terms of fertility and the current methods used to deal with patients that struggle to achieve pregnancy and/or parity.

### 2.2.1 General overview

Under natural conditions, the menstrual cycle governs the female body preparation for the possibility of pregnancy each month. It is regulated by a complex interplay of hormones and involves several stages, lasting under normal circumstances 28 days on average (Fraser et al., 2011). The menstrual cycle is actually comprised of two superposed and interconnected cycles, the uterine and the ovarian one, both divided between before and after ovulation (Fritz and Speroff, 2011). On the uterine level, first there is the menstrual phase, where the lining of the uterus is shed in the form of menstrual blood; and the proliferating phase, where it starts to thicken again. Happening at the same time, on the ovarian level, the follicular phase is happening, where the hypophysis in the brain releases follicle-stimulating hormone (FSH), which causes the available follicles in the ovaries to start growing. One of these follicles will then become dominant and continue to develop until it releases an oocyte during ovulation (Pache et al., 1990; Van Santbrink et al., 1995). As the follicle grows, it produces estrogen (Hillier, Reichert, and Van Hall, 1981), which in time produces the thickening of the lining of the uterus (or endometrium) in preparation for pregnancy (Raine-Fenning et al., 2004).

Ovulation (release of the oocyte from the follicle) happens around the 14th day of a 28-day cycle and lasts for about 24-48 hours (Lenton, Landgren, and Sexton, n.d.). The surge of luteinizing hormone (LH) from the hypophysis triggers the dominant follicle to release the mature oocyte from the ovary (Fritz and Speroff, 2011), which then travels down the fallopian tube towards the uterus (Baerwald, Adams, and Pierson, 2012; Van Santbrink et al., 1995). The mature oocyte is in the meiotic stage metaphase II, and is commonly referred to as MII. Here is when the window of opportunity for conception opens, as if a spermatozoon finds and fertilizes correctly the mature oocyte, an embryo can develop and eventually implant in the receptive endometrium (Küpker, Diedrich, and Edwards, 1998). For that to happen, sexual intercourse is optimal if timed within

two days before ovulation happens (Wilcox, Clarice R. Weinberg, and Donna D. Baird, 1996). Estrogen levels fall after ovulation (Fritz and Speroff, 2011).

After ovulation, on the ovarian level, the corpus luteum (the follicle that contained the mature oocyte) produces progesterone and estrogen to maintain optimal conditions for pregnancy. If pregnancy ends up not happening, the corpus luteum ceases to exist and those hormones levels drop (Fritz and Speroff, 2011; Khan-Dawood et al., 1989). This is called the luteal phase and lasts about 14 days. On the uterine level, the endometrium enters the secretory phase, where the rise of progesterone halts its thickening but makes it receptive to implantation of an embryo. If no such thing happens, progesterone and estrogen levels fall and the cycle starts anew with the menstrual phase (Fritz and Speroff, 2011).

Infertility is defined as the failure to conceive or carry a pregnancy to term despite having regular and unprotected sexual intercourse for at least one year (World Health Organization (WHO), 2018). It is estimated that it affects between 8 to 12% of reproductive-age population in the world (Inhorn and Patrizio, 2014).

Assisted reproductive technologies (ART) are a collection of techniques designed to help individuals or couples that struggle with infertility. These techniques are varied in their grade of invasiveness, from more natural treatments like timed intercourse to more intrusive ones like in vitro fertilization (IVF). A brief definition of the main techniques follows:

- **Timed Intercourse:** This treatment option is the simplest one, and implies the monitoring of the menstrual cycle via ultrasound and hormonal tests in order to time sexual intercourse around ovulation.
- **Intrauterine Insemination (IUI):** This technique also monitors the ovarian cycle via ultrasound and hormonal testing, in order to introduce in the uterus a prepared and concentrated sperm sample around ovulation. The sample can be from either a male partner or a sperm donor. Additionally, the cycle can be natural and just monitored or controlled via a mild ovarian stimulation.
- **In Vitro Fertilization (IVF):** It is the most invasive technique. It involves the surgical aspiration of grown follicles under ultrasound guidance to collect mature oocytes, and collection and preparation of a sperm sample in order to fertilize the oocytes in vitro to obtain embryos. The embryo or embryos considered to be better are transferred back to a prepared uterus in hopes of achieving a pregnancy.

Depending on the cause or combined causes of infertility, expert clinicians will determine which course of action is deemed as preferable. IVF is prescribed whenever IUI has failed or the patient/couple did not even qualify for this first-line procedure due to, for example, low sperm count or fallopian tube obstruction.

In this thesis, the main interest inside ART is IVF, and specifically two subtasks of individual steps: first FSH dose selection for controlled ovarian hyperstimulation (COH), and embryo quantity selection. First, we will review briefly the main steps of IVF, which can also be visualized in Figure 2.4. Next, we will go over current AI applications for each step of IVF, with special detail in the focused interests of this project.

The main steps of an IVF treatment are:

- **Controlled Ovarian Hyperstimulation (COH):** As previously stated in the abridged definition of IVF, retrieval mature oocytes is needed in order to fertilize them in vitro. To retrieve them, first, available follicles in the ovaries need to be stimulated to grow, triggering maturation of the oocytes inside. As stated when reviewing natural conception and the menstrual cycle, follicles grow when FSH levels raise. This process can be monitored with minimal external intervention, which would lead to one dominant follicle to grow, but commonly is supported by administering external FSH. This allows for multiple follicles to grow at once, so leading to collection of more than one oocyte. To administer external FSH safely and guiding stimulation in a controlled way, endogenous FSH, and LH liberation by the hypophysis needs to be suppressed. In order to do that, suppression of the Gonadotropin-releasing hormone (GnRH) secreted by the hypothalamus is needed. That can be achieved either by using agonists or antagonists of GnRH. Once the patient's hormones are down-regulated, exogenous FSH can be administered. After follicles are recruited and grown up to a diameter deemed correct, ovulation can be externally triggered, and the next step scheduled.
- **Ovum Pick Up (OPU):** Once ovulation is pharmacologically triggered but before spontaneous ovulation (that is, around 36 hours after administration of trigger), surgical retrieval of the mature oocytes is scheduled. Often under sedation administered by an anesthesiologist, follicle fluid will be aspirated by a needle guided by vaginal ultrasound. Embryologists will search under microscope for cumulus-oophorus complexes (COCs), which are the combination of granulosa cells that surround the oocyte, and the oocyte itself. Once identified and collected, those COCs will be prepared for fertilization.
- **Sperm preparation:** Sperm samples are usually collected fresh on the same day of the OPU. Collection before that day is also possible, with the consequential freezing of the sample to allow for its use on the day of OPU. The sample, fresh or thawed, is then processed to separate the sperm cells from the seminal fluid. This can be done through various methods, such as density gradient centrifugation, swim-up or microfluidic sperm sorting. These techniques help to select and concentrate the healthiest and most motile sperm for IVF.

- **Fertilization:** Generally speaking, there are two main techniques to fertilize in vitro the collected oocytes. One is *conventional IVF*, where COCs are cultivated overnight with a sample of the prepared sperm. After the needed hours pass, mechanical denudation of the oocyte is performed, discarding the granulosa cells. The oocytes are observed under microscope to determine if they have been correctly fertilized. On the other hand, *intracytoplasmic sperm injection* (ICSI), consists in the manual selection of individual sperm cells and their injection on chemically and mechanically denudated mature oocytes. This process is performed by highly skilled embryologists. The injected oocytes are cultured overnight, and again reviewed under microscope to deem if they have been correctly fertilized, and thus, are able to form apt embryos.
- **Embryo transfer:** Correctly fertilized oocytes, or zygotes, go on to embryo stage when they start cell division. Culture in the laboratory of those embryos since the day of fertilization can span up to seven days. During those days, embryo development is carefully observed by embryologists. Depending on the number of embryos and their quality, transfer back to uterus can happen in days 2-3 of evolution post-fertilization (cell stage) or day 5 (blastocyst stage). The best embryo or embryos will be selected for transfer back to the uterus, while the surplus of apt embryos will be cryopreserved. The number of embryos to be transferred usually ranges from 1 to 3, always keeping in mind that the objective is a single pregnancy, and transferring more than one embryo is only done to increase chances of single pregnancy with reduced risk of multiple pregnancy.
- **Pregnancy test:** After embryo transfer a small period of wait ensues, where if implantation happens a raise of human chorionic gonadotropin (hCG) is expected to happen. It is consistently detectable 15 days after conception, which would be around 9 to 16 days after transfer, depending on the embryo stage when it was transferred. Usually, blood tests are performed to ensure low chances of a false negative.

Although significant strides have been made in the last 40 years, the mean pregnancy rate after an IVF cycle still hovers around 30%, with a 20% chance of delivery (De Geyter et al., 2018). This, of course, even if it betters the chance of pregnancy for patients that need these treatments when compared to natural conception, can be very frustrating for both patients and professionals of the field.

Basic and clinical research is continuously moving forward in hopes of finding more ways of helping these patients by improving treatment success. In parallel, AI has been also developing rapidly, and as aforementioned, it has been explored more and more in multiple healthcare disciplines. It is no exception in the case of ART.

As reviewed just before, an ART treatment, and specially, an IVF cycle, is composed by multiple and often complex steps, where frequently decisions must be made in a personalized way in order

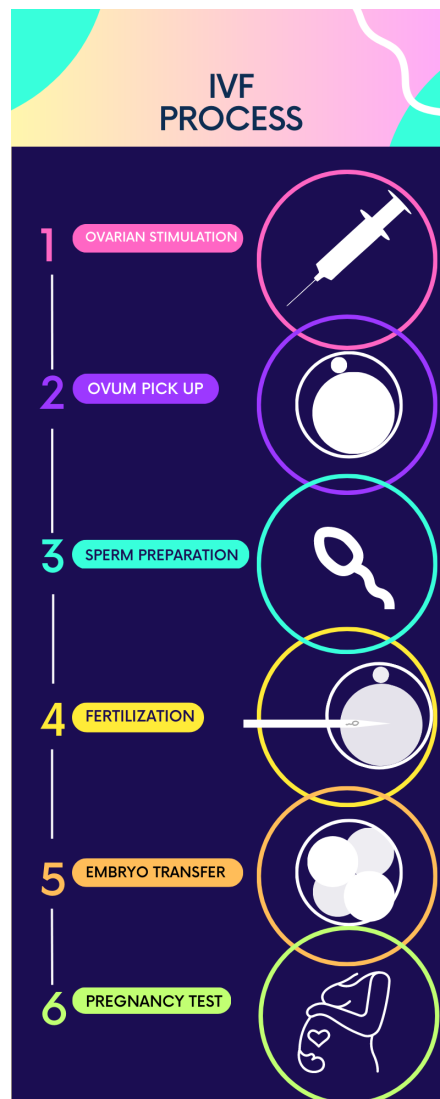


Figure 2.4: Graphical summary of the steps of the IVF treatment. Customized from the original template design published in canva.com by @martaborreguero.

to obtain the best possible outcome. Thus, there are many situations where the application of AI technologies can be of help. The first publication of AI applied to ART dates back to 1997 (Kaufman et al., 1997), but it is not until few years ago that this intersection of disciplines has started to grow exponentially, as it is indicated by the sudden increase in number of AI related presentations in both the American Society of Reproductive Medicine (ASRM) and European Society of Human Reproduction and Embryology (ESHRE) annual meetings of 2018 (Curchoe and Bormann, 2019).

Regardless of this increase in publications, clinical adoption is still slow, and this is directly related to the same general challenges of AI in healthcare. Most critically, professional and patient trust in AI is still tenuous, specially due to the extended use of DL. The general consensus is that

explainability of any model proposed is highly recommended to mitigate this concern (Simopoulou et al., 2018; Curchoe, Malmsten, et al., 2020; Fernandez et al., 2020; Curchoe, 2023). Explainable models may be more accepted simply because the biological logic behind their decisions can be verified. Additionally, early detection of undue bias is necessary to avoid harmful consequences of their use in clinical practice.

One more challenge that concerns professionals is data availability and algorithm standardization. In the field of ART, data isolation is very pronounced, due to interclinic competition, difficulty in the acquisition or absence of data sharing consent by patients, and strict data privacy (Curchoe, Malmsten, et al., 2020). This leads to the emergence of multiple local solutions, few of which are externally validated. In this context, external validation would mean testing the developed model with the data of other clinics, as a means to validate its robustness and reproducibility (Fernandez et al., 2020; Geampana and Perrotta, 2023).

Even if these hurdles are dealt with, there is still some resistance to adopt AI-based solutions due to the opinion that they could substitute human professionals (Simopoulou et al., 2018). However, general opinion of professionals dedicated to AI applications for ART is far from that notion. In fact, AI driven ART is envisioned as a way to enhance human capabilities, helping in the decision-making processes by improving accuracy, speed and reducing variability (Medenica et al., 2022; Curchoe, Malmsten, et al., 2020).

In step with the rising interest and concerns around AI driven ART the need to organize professionals of this intersection has resulted in the celebration of the first international congress on AI in fertility in September 2022. Critical strengths and weaknesses were discussed along with future opportunities and challenges. Its proceedings are thoroughly reviewed in Curchoe, 2023. In the wake of its conclusion, the AI Fertility Society was created with the goal to coordinate international efforts, to create agreements on necessary frameworks and validations, ethical use of AI, publication standards and peer review, collaboration with policymakers, etc. Additionally, a special group interest (SIG) on AI for ART was funded in ASRM at the end of 2022, and effort on the same direction is being carried out for ESHRE.

Next, we will review details of current state-of-the-art on AI solutions for relevant IVF processes, specially for the two key steps pertinent for this thesis. An overview of the contributions described in this chapter can be found in Table 2.1.



ART process	Task	Main Contributions
Treatment prognosis	Male surgery intervention	Ory et al., 2022; Zeadna et al., 2020
	IUI	Ranjbari et al., 2021
	IVF	Choi et al., 2013; Scott M. Nelson et al., 2015
	Number of oocytes needed to obtain a euploid embryo	Esteves, Carvalho, et al., 2019
COH	Selection of starting dose of FSH	La Marca et al., 2012; Ebid et al., 2021; Allegra et al., 2017; Nyboe Andersen et al., 2017 Howles et al., 2006; Olivennes et al., 2015; Fanton, Nutting, Rothman, et al., 2022
	Ovulation trigger date selection	Hariton et al., 2021; Fanton, Nutting, Solano, et al., 2022; Liang et al., 2022
ICSI	Sperm cell selection	Mirroshandel, Ghasemian, and Monji-Azad, 2016; Mirsky et al., 2017 McCallum et al., 2019; Mendizabal-Ruiz et al., 2022
	ICSI automation	Borges et al., 2022; V. S. Jiang, Kartik, et al., 2022
Embryo culture evaluation	Pronuclei detection	Fukunaga et al., 2020
	Morphokinetic events automatic annotation	Danardono et al., 2022; Feyeux et al., 2020
Embryo transfer	Embryo selection	- Embryo-Uterus modelling: Stephen A. Roberts and Stylianou, 2012; Corani et al., 2013; Gianaroli et al., 2013 Hernández-González, Inza, et al., 2018; Hernández-González, Valls, et al., 2022 - Morphokinetics as input: Petersen et al., 2016; Conaghan et al., 2013; Milewski et al., 2017 Campbell et al., 2013; Basile et al., 2014; De Gheselle et al., 2022 - Raw image as input: Cao et al., 2018; Miyagi et al., 2019; Khosravi et al., 2019; Tran et al., 2019 Bormann et al., 2020; VerMilyea et al., 2020; Berntsen et al., 2021 Erich et al., 2022; V. S. Jiang, Kandula, et al., 2023; Ben-Meir et al., 2022 Diakiw et al., 2022
	Selection of the number of embryos	Stephen A. Roberts, 2007; S. A. Roberts, Hirst, et al., 2010 S. A. Roberts, L. McGowan, et al., 2010; Stephen A. Roberts, Linda McGowan, et al., 2011 Lannon et al., 2012; Vaegter et al., 2019; Wen et al., 2022
	Endometrial receptivity	Diaz-Gimeno et al., 2022; Zhang et al., 2021; He et al., 2023

Table 2.1: Main contributions found in the current literature for AI driven solutions in different ART tasks.

## **COH protocol personalization**

An important requisite to the success of an IVF cycle is the availability of a certain number of mature oocytes, usually obtained after COH. Ovarian stimulation, therefore, represents, a key step for IVF success, as failing to ensure an optimal number of MII oocytes will likely hinder the positive outcome of the procedure (a live birth).

As the number of MII oocytes retrieved increases, so does the chance of producing embryos with high pregnancy potential (Drakopoulos et al., 2016; Esteves, Carvalho, et al., 2019), but stimulating a patient excessively leads to an increased risk of ovarian hyperstimulation syndrome (OHSS). As such, a compromise must be reached to retrieve a number of oocytes inside of an optimal range that does not increase chances of OHSS but maintains good pregnancy potential. One such definition of an optimal range of oocytes ranges from 10 to 15 oocytes (Sesh Kamal Sunkara et al., 2011; Steward et al., 2014). Anything outside these values is considered too many or too few. Whenever a patient falls outside the defined range, the risk of an unsuccessful or cancelled cycle increases as well as the occurrence of OHSS. An OHSS risk also implies the need to freeze all the embryos when generated, which increases costs and causes delays in treatment.

Acceptance of an increased risk of OHSS, when properly managed with gonadotrophin releasing hormone agonist trigger, in exchange for a higher number of MII oocytes, is controversial. Sesh Kamal Sunkara et al., 2011 and Steward et al., 2014 found that live birth rates (LBR) in fresh cycles with more than 15 oocytes plateaued or even declined; other investigators (Ji et al., 2013) showed an increased cumulative LBR when the frozen embryo transfers were taken into account. This could benefit patients with specifically advanced maternal age but not patients with polycystic ovary syndrome (PCOS) (Chen et al., 2017).

Essential to all ovarian stimulation protocols is the starting dose of exogenous FSH. This dose should be sufficient to recruit enough FSH responsive follicles but should not be any higher to avoid unsafe effects, i.e. OHSS or decreased oocyte quality Luo et al., 2022. After about 8 days of stimulation, changing the FSH dose does not allow for a significant further recruitment of follicles (Fleming et al., 2006). In other words, if the starting dose of exogenous FSH is inadequate, little can be done to fix its effects on MII yield.

Another relevant decision when personalizing a COH protocol is when to stop stimulation and trigger ovulation. Timing well this intervention will ideally lead to retrieval of a high rate of oocytes grown in the follicles in a mature state, or MII stage (Mohr-Sasson et al., 2020; Permadi et al., 2021). Of course, there are more decisions, like choosing antagonists or agonists of GnRH for the down-regulation of the hypophysis, but there is yet no literature on AI applications for these other decisions.

#### ❖ *AI for FSH dose selection*

In this thesis, we strive to develop a dosing model that finds a first dose of FSH with the best balance of risk-benefit chances for all the diverse population of patients. Chapters 3 to 5 will cover in detail our proposals on the matter. But first, we must understand the current clinical protocol of FSH dosing, and its pharmacometric properties. A comprehensive analysis of FSH pharmacometrics, current standard clinical protocol and published AI solutions can be found in subsection 2.2.2.

#### ❖ *AI for trigger time selection*

In recent years, there has been some exploration of AI solutions to select the right time to trigger ovulation, with few publications in this direction, but the ones available are quite interesting and diverse.

Hariton et al., 2021 proposed the use of causal inference where a binary treatment (trigger or wait) was assessed in its effect on the outcome measured as number of correctly fertilized oocytes (2PNs) and usable blastocysts. Fanton, Nutting, Solano, et al., 2022 used linear regression to generate two predictive models (triggered today, triggered tomorrow) and used PSM (as with their FSH starting dose) to review expected improvements retrospectively. Lastly, Liang et al., 2022 presents an AI tool to segment the images obtained by a 3D ultrasound scan, obtaining the volumes of the follicles, and using those volume values to better tailor the trigger date. All three show expected improvements, but are not tested prospectively yet.

### **Embryo transfer personalization**

#### ❖ *Embryo selection*

Selecting from the patient's cohort the embryo with the highest chances of pregnancy is of great importance for cycle success. If selected accurately, the patient may get pregnant at the first embryo transfer. If not, the patient will need to undergo a new transfer attempt, implying more time to pregnancy and higher economical and psychological burden.

Since the dawn of IVF, embryos generated in the laboratory have been ranked by their morphological traits observed once a day under the microscope. The Istanbul consensus (Balaban et al., 2011) brought standardization to many international morphology grading systems. Even-though scoring methods based in morphology have been consistently proved to be positively correlated to pregnancy rates (Balaban et al., 2011; Gardner et al., 2000), pregnancy rates after an IVF cycle remain around 30%. This could be related to intra and inter-operator variability when using this scoring method (Martínez-Granados et al., 2018; Storr et al., 2017; Arce, Ziebe, et al., 2006; Paternot et al., 2011), and to the limited information that can be obtained through a morphological evaluation on the implantation potential of any embryo.

Some of the first AI-driven solutions to optimize the selection of the embryo with the highest expected chance of pregnancy were based on bayesian networks, introducing as predictor variables information on the morphology of the embryos, cycle characteristics and/or uterine features (Speirs et al., 1983; H. Zhou and C R Weinberg, 1998; Stephen A. Roberts, 2007; Morales, Bengoetxea, and Larrañaga, 2008). This has been referred to as the embryo-uterus (EU) model. More recent publications use the same strategy (Stephen A. Roberts and Stylianou, 2012; Corani et al., 2013; Gianaroli et al., 2013; Hernández-González, Inza, et al., 2018; Hernández-González, Valls, et al., 2022), as bayesian networks provide the advantage of being able to deal with partially labelled datasets (we don't know the pregnancy results for all embryos as not all are transferred), and are easily explainable due to their structure being formulated as a variable inter-dependency graph. However, these models could not provide a significant improvement over the morphology scoring system.

Morphological scoring systems still get some update following literature (Cuevas Saiz et al., 2018), but new scoring methods appeared since the introduction in the laboratory of time-lapse system (TLS) technologies. These new incubators allow for a continuous visual evaluation of the developing embryos without removing them from temperature and pH controlled conditions. Subsequently, new information on embryo development has become available, like morphokinetics, which refers to the specific timing for certain embryo development events (like cell division or compaction). Models to exploit this new information in order to select the best embryo of the cohort were developed (Petersen et al., 2016; Milewski et al., 2017; Conaghan et al., 2013), but the posterior meta-analysis by S. Armstrong et al., 2018 reviewed the RCTs performed with these models, showing a lack of significance in pregnancy results versus control cases. Some of the reasons for the results obtained brought forward by the authors are that manual morphokinetics annotations suffer of intra and inter-operator variability (as with morphologic scoring), and the lack of well designed RCTs. Aneuploidy detection was also used as end-point in some models that employed morphokinetic data as predictor variables (Campbell et al., 2013; Basile et al., 2014; De Gheselle et al., 2022).

New solutions were explored by applying DL to the bulk of images obtained thanks to the TLS, with very promising results in prediction of embryo quality, viability or pregnancy results (Cao et al., 2018; Miyagi et al., 2019; Khosravi et al., 2019; Tran et al., 2019; Bormann et al., 2020; VerMilyea et al., 2020; Berntsen et al., 2021; Erlich et al., 2022). The same methodologies are being applied to predict chances of euploidy of the analyzed embryos (V. S. Jiang, Kandula, et al., 2023; Ben-Meir et al., 2022; Barnes et al., 2023; Diakiw et al., 2022). These solutions, though very interesting, lack still prospective validation, and suffer from the curse of the black box. There is, though, awareness of the lack of explainability in the field, as some researchers strive to make the models interpretable (Hickman et al., 2022).

#### ❖ *Selection of number of embryos for uterine transfer*

For many years, it has been common practice to transfer two embryos simultaneously to the uterus in order to improve success rates and increase the likelihood of achieving a pregnancy compared to Single Embryo Transfer (SET). (Kamath et al., 2020). Consequently, Double Embryo Transfer (DET) stands as a 54.5% of all embryo transfers. The unwanted consequences of this high DET prevalence is an increased obstetrical risk due to the increment of instances of multiple pregnancy. Compared to single pregnancies, twin births are four times riskier (Crosignani et al., 2000). As such, a multiple pregnancy is an undesired outcome of ART cycles. An equilibrium must be achieved between the raised chances of pregnancy of DET, and the almost null incidence of multiple pregnancy of SET, as particular patient factors (economical, psychological or individual success prognosis) can modulate the selection of number of embryos to be transferred.

As this problem is of special interest in this thesis, a close-up on it and the current state of literature of AI applications for it can be found in subsection 2.2.2.

#### ❖ *Endometrial receptivity detection*

Successful implantation, and hence, the start of a pregnancy, does not depend uniquely on the embryo. The uterus, specially the endometrium, is the other main player in this process. The human endometrium is a constantly changing tissue that becomes receptive to the implantation of a blastocyst in response to steroid hormones during a brief time frame known as the window of implantation (WOI) (Harper, 1992), which begins around days 19-21 of the menstrual cycle (Wilcox, Donna Day Baird, and Clarice R. Weinberg, 1999). The endometrium exhibits a receptive state during the WOI, which allows for implantation to happen, characterized by various processes such as adhesion, invasion, survival, growth, differentiation, decidualization, and immunomodulation (Carson et al., 2000).

Identifying the inter-patient variable WOI is still a topic of research, as inadvertently transferring an embryo on a misaligned WOI timing can cause a failure of implantation (Haouzi et al., 2012; Ruiz-Alonso et al., 2013). Tools powered by AI are already present in the literature, and some in the market, that use endometrium transcriptomic analysis results in order to identify the timing of WOI for individual patients (Diaz-Gimeno et al., 2022; Zhang et al., 2021; He et al., 2023).

### **AI driven sperm selection**

Frequently in IVF, the focus of attention is centered in the oocyte or the embryo, leaving the sperm cell in a undeserved second plane. This has been the case too for AI driven solutions in IVF, as much of the effort is invested in embryo-related solutions. There has been some relevant advances though, with early implementations delving good results but not being yet able to integrated in the laboratory workflow (Mirroshandel, Ghasemian, and Monji-Azad, 2016; Mirsky et al., 2017;

McCallum et al., 2019). A recent publication solved that by making their software compatible with micro-manipulators used for sperm selection during ICSI (Mendizabal-Ruiz et al., 2022), delving a tool that enables real-time AI supported selection of the best spermatozoa.

## **Other AI driven solutions for ART**

### **❖ *Automation***

In recent years, a new trend in automation of processes in the IVF laboratory has emerged in the scientific literature. We had previously stated that manual morphokinetic annotation suffers from intra and inter-operator variability, which "contaminates" the predictive models with the subjectivity of the operators. One solution is the automation of those processes (Feyeux et al., 2020; Fukunaga et al., 2020; Danardon et al., 2022).

Further research is being performed in the automation of more complex processes like ICSI with promising results (Borges et al., 2022; V. S. Jiang, Kartik, et al., 2022). In the near future we could be witnesses to a near-to-fully automated IVF laboratory under the supervision of human embryologists.

Even if AI-driven solutions do not seek to substitute human professionals, certainly there will be a shift on the tasks embryologists and clinicians will do after enough AI-driven automations are deployed in the clinical workflow.

Automation helps with standardization of processes, and frees time of the professionals previously sequestered by those technical procedures. This shift in tasks and re-acquisition of free time will usher a new type of professional, on the words of Daniella Gilboa, a *computational embryologist* (Curchoe, 2023). This new type of embryologist will be able to use their experience for high order data analysis, research, quality control and mentoring.

### **❖ *Cycle success prediction***

Research is very much focused on the development of deep knowledge on all reproductive processes, and specifically, on using this knowledge to raise the success chances for all patients. This leads to the focus of AI related investigation to focus on interventional models, or in other words, in models that optimize specific steps of the treatment by adjusting it.

But in the day to day practice, patients, who more often than not rely on professionals for those decisions and rarely weigh-in, ask routinely about one topic: expectation of success for their individual case. Usually, clinicians give estimated success rates stratified by age range, computed for historical records (per center or from national records). Recently, studies have been published covering the matter. From success rates estimated before commencing an IVF treatment (Choi et al., 2013; Scott M. Nelson et al., 2015), and IUI (Ranjbari et al., 2021).

### ❖ *Other applications*

There are other applications of significant clinical relevance that are also highly compelling. For example, the proposal by Esteves, Carvalho, et al., 2019, where a calculator is able to compute how many mature oocytes are required to obtain at least one euploid blastocyst. This type of proposal could go hand in hand with a FSH starting model, as depending on the number of oocytes predicted, the clinician would want to push the ovaries differently, and thus, prescribe a different FSH dose.

Other significant publications are those by Ory et al., 2022 and Zeadna et al., 2020, where an ML model predicts the results of surgical interventions for the male partner. Specifically, of varicocele repair and testicular biopsy respectively. This is clinically compelling, as achieving good prediction results can enhance the decision power for clinicians faced with patients that may need it, and avoid futile surgical interventions.

## **2.2.2 Dose recommendation in IVF**

In this thesis the focus lies in the optimization of dose selection policies in IVF processes, training models that abide by the available clinical knowledge. This subsection details the context for the two main processes of interest: selection of the first dose of FSH for COH and selection of the number of embryos for transfer.

### **Selection of the first FSH dose for COH**

Next, a close zoom into FSH dosing policies for IVF treatments is made. In order to understand fully the background of the problem, first, FSH pharmacometrics will be described. Next, current standard dosing policies for FSH will be explained. Lastly, the state-of-the-art of AI driven solutions for FSH dosings will be examined.

### ❖ *FSH pharmacometrics*

The gonadotrophs found in the anterior pituitary gland (hypophysis) are responsible for the synthesis and secretion of FSH when stimulated by the hypothalamic GnRH (Richards, 1980). FSH's  $\beta$ -subunit is responsible for its functional specificity, ensuring that it interacts exclusively with the FSH receptor (FSHR) (D. T. Armstrong and Dorrington, 1976). FSHRs are located in the granulosa cells of the follicles in the ovaries. As mentioned, sigmoid equations often describe closely dose-response relationships where protein-receptors interactions are involved. FSH is no exception, as both literature and clinical experience show evidences that point in that direction.

This is clearly reflected in the results of the pharmacometrics studies by Porchet, Le Cotonnec, and Loumaye, 1994 and by Arce, Klein, and Erichsen, 2016 where sigmoid functions of the type

E-max were used to fit the PD portion and described adequately the study data in both cases. Abd-Elaziz et al., 2017 did not explicitly use a sigmoid function to fit PD data, but did report a positive relationship between FSH serum levels and follicular growth.

In the clinical literature, there is evidence of this same positive relationship, in other words: generally, the more dose of FSH administered, the more follicles are recruited to grow, and hence, the more oocytes that can be retrieved (Abbara et al., 2019; Lensen et al., 2018) provided that spontaneous ovulation is prevented. This pattern replicates almost always across different types of patients, from expected poor responders to expected high responders. Expected poor responders are defined as those patients that have  $< 5$  AFC and AMH levels of  $< 1.2$  ng/ml (Conforti et al., 2019). Expected high responders can be defined as those that have  $> 24$  AFC and levels of AMH of  $> 3.4$  ng/ml (ASRM, 2021; Sun et al., 2021).

A sigmoid function also describes a saturation phenomenon, where from a certain threshold of dose, the clinically relevant response to the drug remains unvaried. Clinically, this saturation is described specially in low responders, as seen also in the meta-analysis by Lensen et al., 2018, where up to 300 UI of exogenous FSH a positive relationship is observed, but no statistical difference is observed over this threshold. From a physiological standpoint, this is also plausible, as a patient with fewer available follicles in the ovaries, where the FSH receptors are located, would reach saturation sooner with the same dose of FSH than a patient with more follicles ready to grow, and thus, with more FSH receptors available.

#### ❖ *Current standard clinical protocol for FSH dosing*

Although this sigmoid-like relationship is quite clear both from the pharmacometrics and from the clinical point of view, there are many bio-markers that modulate this relationship. Consequently, not all patients react equally to the same doses of FSH.

In clinical practice, the choice of the FSH starting dose is mostly based on the patient's characteristics, i.e. age, body mass index (BMI) or ovarian reserve and clinical characteristics, i.e. past gravidity and parity. Among these bio-markers, some modulate the reactivity to FSH downwards, like age (Shahrokh Tehraninezhad et al., 2016; Wilkosz et al., 2014; Amanvermez and Tosun, 2015), BMI (Imterat et al., 2019), basal FSH (Abdalla and Thum, 2004), etc. On the other hand, higher levels of anti-Müllerian hormone (AMH) and antral follicle count (AFC) regulate upwards the reactivity to FSH (Hansen et al., 2011; Anderson, S. M. Nelson, and Wallace, 2012). Additionally, several of these bio-markers are correlated. For instance, it is well described that when age increases, ovarian reserve decreases. In more detail, levels of AMH and AFC decrease, and basal FSH increases (Hansen et al., 2011; Anderson, S. M. Nelson, and Wallace, 2012; Shahrokh Tehraninezhad et al., 2016; Amanvermez and Tosun, 2015; Steiner et al., 2017). This is directly related to and describes the decline in the number of available follicles in aging ovaries. Regardless, this decline is not consistent for all women, and a combination of these bio-markers is usually



used to determine the preferred starting dose of FSH to reach an optimal outcome.

Sometimes, in spite of the careful evaluation of multiple bio-markers, ovarian stimulation leads to unexpected and widely different results even among apparently similar patients, resulting in either too many or too few oocytes collected. Furthermore, even if the MII oocytes retrieved are in the expected number range for the biomarkers analyzed, they may still be of insufficient quality to achieve success, as only 30–40% of microinjected oocytes develop to blastocyst (Maggiulli et al., 2020; Vaiarelli et al., 2020), and around 11% to an euploid blastocyst (Chamayou et al., 2017). Here is important to remember that even-though in the dose-response relationship at hand the outcome is number of oocytes retrieved, the main desired outcome of an IVF cycle is healthy baby at home. Thus, the importance of striving to reach a number as close as possible to an optimal range for all patients.

Experienced clinicians utilize past cycles outcomes in order to predict and avoid cases of unexpected response, but they are generally unable to detect these deviations from normality if no previous cycle's results are available. This is specially true whenever FSHR polymorphisms are involved, as they may hinder FSH sensitivity of patients affected, needing more FSH dose to reach the same outcomes (Lledo et al., 2014). Although significant differences have been described in number of oocytes retrieved for patients carrying known polymorphisms of FSHR, there is discussion on the clinical relevance of testing for these genetic variants before treatment (Nikolaos P. Polyzos et al., 2021; Neves et al., 2022).

Continued research has led to the publication of guidelines for COH (The ESHRE Guideline Group on Ovarian Stimulation et al., 2020), but regarding FSH dose only cover low responders. General consensus on protocols to prescribe first dose of FSH is not fully consolidated, and efforts are being centered in improving it (Barrenetxea et al., 2023).

#### ❖ *AI for FSH dosing*

So far, some machine learning models have been developed to encapsulate the medical experience reflected in historical data to try to automate that decision. Two separate nomograms based on patient age, AMH or AFC and basal FSH levels have been developed for this task (La Marca et al., 2012; Ebid et al., 2021). The nomogram by La Marca et al., 2012 was tested prospectively (Allegra et al., 2017), reporting an increase in the number of patients with an optimal range of MII oocytes retrieved, and a decrease in the number in patients with lower response in those using the nomogram. These two nomograms did not include patients older than 40 years or those with irregular cycles, including patients with PCOS. In an RCT for another model developed specifically for individualized dosage of FSH delta (Nyboe Andersen et al., 2017), no differences in pregnancy rate were observed.

Additionally, the CONSORT model, based on multivariate regression (Howles et al., 2006) predicted overall lower starting doses compared with those prescribed by clinicians in normo-ovulatory

patients (Naether, Tandler-Schneider, and Bilger, 2015; Pouly et al., 2015). CONSORT was also tested by RCT (Olivennes et al., 2015), showing that the model was able to reduce the risk of OHSS in patients while maintaining comparable pregnancy rates compared with the clinician-chosen dose, despite a reduction in the number of retrieved oocytes.

A recent study (Fanton, Nutting, Rothman, et al., 2022) published after our first two iterations (covered in Chapter 3) does cover all types of patients, and introduces the concept of computing individual dose-response curves. It does so by searching the 100 most similar patients using *k*-nearest neighbors (KNN) and fitting a constrained second-order polynomial to data on the number of MII retrieved ( $y$ ) and the FSH starting dose administered ( $x$ ). Mostly flat curves were deemed as non-responsive to FSH, which was the case for 30% of the cases analyzed. For dose-responsive curves, the optimal starting dose was identified as the one where the curve showed a peak of MII. Using propensity score matching (PSM) to pair similar patients with different doses, they concluded that patients that received the optimal dose predicted by their model versus those who didn't obtained better results. Although the design is indeed very interesting, it is surprising that so few patients analyzed are predicted to be dose-responsive. This could be due to a low variability of doses prescribed for patients deemed flat-responsive, probably concentrated in doses where saturation of FSH receptors has already been reached for these patients. Additionally, the curves used to fit the dose-response relationship do not fit well with the known pharmacometrics of FSH. Second-order polynomial curves do not reflect the positive monotonicity (as dose increases, outcome also increases or remains the same, but never decreases), nor the saturation properties of sigmoid curves, which are deemed to be the closest fit for the FSH-oocytes relationship.

A common characteristic of these proposals and our own (exposed in Chapters 3 and 4) is that they are very transparent and explainable. Using linear, sigmoid, or curve functions, it is easy to interpret how the trained models work. We mentioned before this was capital for healthcare related AI, and it is clear that it has been prioritized by all research teams in this specific task. This will be a returning theme during this thesis, as it has been a paramount pillar during this whole work.

### **Selection of number of embryos for uterine transfer**

To mitigate low success rates, the transfer of two embryos simultaneously to the uterus has been the standard during many years. This certainly increases the chances of achieving a pregnancy versus Single Embryo Transfer (SET) (Kamath et al., 2020). 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 (Crosignani et al., 2000). 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, RCTs have consistently shown that SET provides lower pregnancy rates than DET, but they do so with a much lower twin rate (about 1%, mostly due to embryo splitting and monozygotic -identical- twins forming). 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 (Kamath et al., 2020). This would, from a strictly clinical point view, lead to an easy solution, which would be to always perform 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 worse morphology is an indicator of lower development potential and higher aneuploidy rates (Hardarson et al., 2003). In these cases, DET is used as a strategy to allow for higher pregnancy rates in poor prognosis treatments, assuming that the risk of multiple pregnancy should be lower as at least one of the two embryos transferred has a low chance 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) (Glujovsky et al., 2016). 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 (Grøndahl et al., 2017) 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 recommendation based on literature and firsthand 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 to a healthy baby. They participate actively in making the final decision of how many embryos will be 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 miscarried pregnancies, etc.), the economic pressure of the treatments and the information that they receive and/or understand (Abd-Elaziz et al., 2017).

Few studies modelling the number and quality of the embryos to transfer have been carried out in this regard, but the ones that did give us some very interesting insight. One proposal is based on the EU model presented in Stephen A. Roberts, 2007 and later refined iterations (S. A. Roberts, Hirst, et al., 2010; S. A. Roberts, L. McGowan, et al., 2010). It proposes modelling by logistic

regression separately the embryo and the uterine components, combining both probabilities for the final live birth outcome. The final model (composed of both sub-models) was fitted by direct maximization of the resultant observed data likelihood.

In the following publication (Stephen A. Roberts, Linda McGowan, et al., 2011), results are presented on its use for prediction of different scenarios (SET or DET) for each analyzed patient. This allowed them not only to validate the predictive power of the model, but simulate counter-factual scenarios and analyze the predicted results. Simulations results showed that SET had approximately one third less LBR in comparison with DET in a per transfer level. Accurate selection of patients for SET can ameliorate this difference. When considering full cycles (i.e. the results after the transfer of all available embryos from one ovarian stimulation), simulations showed a likely increase in cumulative LBR by using the SET strategy. No prospective validations with this model have been published to date.

In the very thorough report on the theme performed in S. A. Roberts, L. McGowan, et al., 2010, it also recounts construction of nested pregnancy (P) and multiple pregnancy (MP) models using first UK national reports. The nested component relates to how the MP model is constructed solely from data of patients with a positive result of pregnancy, hence, it is used only if the P model predicts a positive result. Results show an AUC of 0.60 for the first model and of 0.66 for the second. Further research includes information from multiple private centers with more predictor variables in the training of the models, resulting in slightly better AUC scores. Interestingly, authors point out that these models, due to their modest prediction powers, are not suitable for individual prediction, but are useful to predict population tendencies.

Lannon et al., 2012 proposes a boosted tree model trained only on patients that got a DET and achieved a pregnancy, a methodology of inclusion of cases similar to that of the logistic regression models by S. A. Roberts, L. McGowan, et al., 2010. To validate the model, it is compared to a similar model developed to only take into account the age of the patient, and to baseline prediction (that is, predicting MP by assigning always the general MP rate in their train dataset). Results showed improvement of predictive power versus both controls, with an AUC of 0.632. This proposed methodology, though, does not allow for comparison between SET and DET scenarios for the same patient.

Another relevant study created independent models: one for P prediction and its pair model for MP, both trained exclusively on DET cycles; and another one for P prediction trained only in SET cycles, getting AUCs between 0.64 and 0.75 (Vaegter et al., 2019). This is the only methodology that presents prospective interventional validation, where whenever a patient had predicted >15% chance of MP after DET was selected for SET. Significantly higher LBR and cumulative LBR (CLBR) were observed in the population where the model was used, versus retrospective data where the model was not used. It is pertinent to point here that the models were constructed with data from embryos up to the second day of development post fertilization.

Lastly, Wen et al., 2022 presented the same methodology of nested P and MP predictive models, determining that XGBoost algorithms achieve best results, with respective AUCs being 0.787 and 0.732. Compared to previous study, even if the performance is better, the size of the training and validations datasets are reduced.

All these studies only report model performance by means of scores like AUC or accuracy, but do not explicitly check compliance of the models with general clinical knowledge. Derived from that knowledge, a set of constraints can be defined, and the models tested against them. One such exploration of this problem is covered in Chapter 6, and a tentative approximation of it as a dose-response problem is described in Chapter 7.

## **2.3 RCT trials for AI solutions in ART**

As reiterated elsewhere, AI-driven solutions in general, but specially in healthcare must be applied with care. Proper evaluation of the behaviour of the models needs to be performed in order to avoid harm to patients. In clinical fields the RCT is the gold standard to assess the efficacy of a treatment. In the ART discipline specifically, experts like Cristina Hickman, PhD lead the discussion around when is an RCT really necessary to prove efficacy of developed tools (Curchoe, 2023).

There are varying levels of autonomy in AI technologies, and each level of autonomy carries its own degree of risk. From assistive AI (informative models, clinical decision support), to conditional automation, and up to full automation AI. However, regulation-wise, often a "one size fits all" approach is adopted and all AI systems are subject to the same level of scrutiny. As a result, clinical decision support models are often cautiously classified by local authorities as medical devices of type IIb (under Regulation (EU) 2017/745), requiring them to conduct clinical trials in order to obtain the Conformité Européenne (CE) marking before clinical implementation. Only if regulators deem the model as a medical device of type I, can the developers auto-certify themselves and not report the results of an RCT in order to implement the tool clinically.

This approach may be driven by concerns around the potential risks associated with AI systems, as regulators consider that even low-risk assistive AI systems can have significant impacts on human lives and decision-making. Additionally, it may be difficult for them to determine the precise level of autonomy and risk associated with each AI system. That is why collaboration between regulators and experts of the field is paramount to protect the safety of patients while simultaneously favoring the progress of research and its translation to technical solutions.

Regardless, nowadays these trials are still being required for many AI developments. Given that RCTs are very expensive and time-intensive, there are not many models designed for ART processes that have an RCT published on their efficacy.

In the case of FSH dosing, the nomogram by La Marca et al. (2012) was tested prospectively in an RCT (Allegra et al., 2017), describing an increased number of patients falling into an optimal range of MII, and a lower number of patients with lower outcome for those participants in the intervention arm. The multivariate regression developed by Howles et al., 2006 was tested also (Olivennes et al., 2015), results showing that the model was able to reduce the risk of OHSS in patients while maintaining pregnancy rates, albeit an observed reduction in the number of recovered oocytes. Lastly, another randomized trial was performed for a model developed specifically for FSH delta (Nyboe Andersen et al., 2017). Results described significantly more patients with target response (8–14 oocytes), less excessive responses ( $\geq 15$  oocytes in the high AMH stratum), and less poor responses ( $< 4$  oocytes in the low AMH tier), always comparing to a non-personalized-dosing control group. Nevertheless, no differences in pregnancy and live birth rates were observed.

Concerning embryo selection, S. Armstrong et al., 2018 analyzed the published RCTs on models developed to support the decision on which embryo to transfer first by training them on morphokinetic data. Results showed that there was no evidence that the use of TLSs alone, or TLSs in combination with predictive models improved pregnancy or live birth. Currently no RCTs are available on the efficacy of pregnancy or aneuploidy prediction models based on image training via deep learning to the best of our knowledge.

Most probably additional RCTs are underway for multiple AI-driven applications for ART, but the current requirement for this type of trial in order to implement the developed solutions is one of the main barriers hindering clinical adoption.

## 2.4 Conclusions

In this chapter, we have reviewed the many possible applications of AI in healthcare in general, and in ART and IVF specifically, including its main pitfalls and strategies to overcome them. IVF has many steps where AI can be of help to optimize results. In this thesis, the main point of interest is the optimization of dosing policies in the IVF treatment, with a dominant focus on selection of the first FSH dose for COH protocols. After evaluation of the current literature, it is clear that there is a lack of AI-driven dosing models that include all patient population without restrictions, and at the same time, upholds pharmacometrics and clinical evidence on its dose-response relationship. In this text we will cover 3 increasingly accurate iterations of dosing models for FSH, a proposed protocol for an RCT that would test its accuracy, and an initial exploration of the selection of number of embryos to transfer.

## Chapter 3

# FSH dosing policy optimization

In this chapter, we will present an initial answer to the question

**Q1:** “*Is it possible to improve clinical FSH dosing policy for Controlled Ovarian Hyperstimulation using only historical data?*”,

and its sub-question

**Q1a:** “*How can we analyze a dosing model’s performance before clinical intervention?*”.

The chapter will be divided in three main sections:

- The first section describes an initial solution (an extended version of “Núria Correa, Flavia Rodríguez, et al. (2021). “P-637 Development and validation of an Artificial Intelligence algorithm that matches a clinician ability to select the best follitropin dose for ovarian stimulation”. In: *Human Reproduction* 36.Supplement.1. deab130.636. DOI: 10.1093/humrep/deab130.636”)
- The second section details a second iteration of the solution (a slightly modified version of “Núria Correa, Jesús Cerquides, Josep Lluís Arcos, and Rita Vassena (2022). “Supporting first FSH dosage for ovarian stimulation with machine learning”. In: *Reproductive BioMedicine Online* 45.5, pp. 1039–1045. DOI: 10.1016/j.rbmo.2022.06.010” and presented as “Núria Correa, Jesús Cerquides, Amelia Rodríguez-Aranda, et al. (2022). “379/427 Acompañamiento en la selección de la dosis de FSH para estimulación ovárica mediante machine learning.” In: *33º Congreso Nacional Sociedad Española de Fertilidad*. Oral communication. Sociedad Española de Fertilidad (SEF). Bilbao”)
- The third section is focused on common learnings and conclusions.

### 3.1 Background

As a brief reminder, at the time of writing, literature on AI for the optimization of the first FSH dose for COH included the nomograms by La Marca et al., 2012 and Ebid et al., 2021, a multivariate regression model (Howles et al., 2006) and a model developed uniquely for FSH delta (Nyboe Andersen et al., 2017). Some of these models were validated via RCT with good results (Allegra et al., 2017; Nyboe Andersen et al., 2017; Olivennes et al., 2015). However, patients older than 40 and/or with irregular menstrual cycles were excluded. Those models then, do not cover the complete range of patients, further, they are not optimized for patients where a correct dose selection is even more critical. The solutions presented in this chapter do not exclude older patients nor patients with irregular cycles.

With the exception of Nyboe Andersen et al., 2017 and Howles et al., 2006 (where data from pharmaceutical Phase II-IV studies was available for the authors), observational data is the common source of information available for researchers. Prospective data is, in general, challenging to obtain when not already available. Additionally, and as mentioned in Chapter 2, observational datasets tend to suffer from confounding, requiring the implementation of measures to address it. Methods to do that, like causal inference, need data sufficiently varied and complete, which is very hard to comply with in clinical settings. Finally, up to the moment of these investigations, models presented in the literature reported performance solely based on scores that evaluate the fit of the model. Scores like concordance probability index (C-index), that measures the association between predicted doses and actual doses; Akaike information criterion (AIC), that quantifies how well the model fits the data it is generated from; or Pearson's correlation. Evaluating performance like this is informative, but does not disclose information on the quality of the recommendations made by the model, especially and more critically when they differ from the clinical ones. This is very relevant in this setting, where models need to be validated by an RCT before clinical implementation. It would be beneficial to devise a novel methodology for assessing dosing models pre-clinically, as the investment of time and resources in such trials is considerable. This would enable more effective screening of potential candidates, and enhanced safety for participants of the prospective trials.

Hence, in this chapter, the challenges addressed are the following:

- Creating an unconfounded FSH dose model with observational data.
- Creating a pre-clinical scoring method of doses proposed by the constructed model.

Additionally, two relevant requirements for the FSH dosing model were set:

1. Obtaining 10 to 15 MII was considered the objective of the selection of the first dose of FSH.



In agreement with published research (Sesh Kamal Sunkara et al., 2011; Steward et al., 2014; N. P. Polyzos and S. K. Sunkara, 2015), the aim of the study was to predict the initial dose of FSH to achieve a number of MII as close as possible to an optimal range of oocytes. The range 10–15 was considered desirable, the range four to nine suboptimal and MII lower than four or above 15 not desirable (see Figure 3.1). Given patient characteristics and limitations in the maximum dose of FSH allowed to be administered, not every patient is considered able, a priori, to reach the desired goal. The number of MII was selected as an end point due to its closeness in time and association with the intervention, while it maintains clinical relevance (a recognized association exists between number of MII and chances of pregnancy and live birth). Live birth rate (LBR) and clinical pregnancy rate (CPR) were considered initially in the building of the model but were too distant in time from the intervention for any model to be able to predict accurately the effect of a specific treatment using, as in the present study, only the information from participants available at the start of treatment.



Figure 3.1: Outcome ranges expressed in number of oocytes considered non-desirable, suboptimal and optimal.

2. The maximum recommended dose of FSH allowed to the model was 300 IU.

As per ESHRE guidelines (OS Guideline Development Group, 2019) based in literature (Harrison et al., 2001; Bastu et al., 2016), starting doses over 300 IU are not recommended due to the absence of significant advantages observed in prospective trials with poor responders.

### 3.1.1 Patient population

Data from a total of 2713 first IVF cycles, from January 2011 to September 2019, registered in five private centers, were used to develop the model. All five centers operated under similar quality-control protocols, but choice of stimulation and modifications to standard protocols were left to each clinician. Natural cycles and cycles in which FSH doses were not expressed in IU/ml were excluded. The inclusion of first cycles aimed to prevent bias caused by unrecorded clinician knowledge (such as FSH dosage and results of previous cycles). An additional 774 cycles between January 2020 and May 2021 were used for prospective validation of the model. Three categories of data were collected as variables. First, the input data, composed of age, BMI, proven fertility

(Y/N) and ovarian reserve markers AMH and AFC; second, the intervention, namely the first dose of FSH prescribed by the clinician; and third, result data expressed as number of MII collected after stimulation (see Table 3.1). Throughout the study, only cases with complete data on all the variables were included. Cycles from both the development and validation databases corresponded to women aged  $37.9 \pm 4.6$  years (18–50 years), with a BMI of  $23.6 \pm 4.2$  kg/m<sup>2</sup>, AMH of  $2.4 \pm 2.3$  an average number of MII obtained  $6.8 \pm 5.4$ .

	<b>Development database (N=2713)</b>	<b>Validation database (N=774)</b>	<b>p-value</b>
<b>Age (years)</b>	$37.8 \pm 4.6$	$38.3 \pm 4.4$	$<0.05^*$
<b>AMH (ng/ml)</b>	$2.4 \pm 2.3$	$2.2 \pm 2.2$	$<0.05^*$
<b>AFC</b>	$11.2 \pm 7.3$	$11.3 \pm 8.5$	0.29
<b>BMI (kg/m<sup>2</sup>)</b>	$23.7 \pm 4.2$	$23.2 \pm 3.9$	$<0.05^*$
<b>Number of MII retrieved</b>	$6.8 \pm 5.1$	$6.7 \pm 6.5$	$<0.05^*$
<b>Proven female fertility</b>	12.4%	10.1%	0.067

Table 3.1: Patient characteristics in the two databases used in the study. Values are expressed as average and SD. Variables were compared using Mann-Whitney U test. For proportions a 2-sample z-test was conducted. A p-value of  $<0.05$  was decided as significant.

## 3.2 First iteration

In order to define whether an FSH dosing model is better than the current clinical practice before any prospective intervention, first, a method to evaluate recommended doses is needed. The next section includes the description of an evaluation methodology tailored for starting FSH doses.

### 3.2.1 Development of a performance score

A score function was designed to compare individual recommendations made by any FSH dosing model with the prescriptions made by the clinicians. Given any FSH prescription, paired to its resulting ground truth MII outcome, the function assigns a performance score value, or  $\varphi$  for a hypothetical recommended dose:

$$\varphi_i = f(y_i, y^*, \hat{d}_i, d_i) \quad (3.1)$$

Where  $y_i$  is the ground truth outcome,  $y^*$  the desired outcome range,  $\hat{d}_i$  the dose to be evaluated, and  $d_i$  the real prescribed dose. The score value ( $\varphi$ ) can span from  $-1$  (the recommended dose is considered too low) to  $1$  (too high),  $0$  being the best possible value (the dose recommended is considered appropriate). FSH doses ( $d_i$  and  $\hat{d}_i$ ) are categorized in four ordinal ranks (100 to 150,

151 to 200, 201 to 250 and 251 to 300) to create the score function. Here, 300 IU is deemed as the maximum dose allowed, as previously stated. The score function is designed to consider appropriate those dose ranks where the outcome is not in the optimal range, but the dose rank is not improvable. For example, it does not penalize a dose rank whenever a poor responder gets recommended and/or prescribed the highest one but has yielded few to none oocytes, as there is little more that is possible to achieve, physiologically, from a poor responder.

The score function also allows clinical prescriptions to be assessed, which can be done by setting the recommended dose equal to the clinically prescribed dose. In doing so, the function evaluates how close the MII outcome is from the optimal range (10 to 15), and if there is any room for improving the dose rank (penalizing prescribed doses that did have margin to improve outcome).

In order to assign a  $\varphi$  value for every possible combination of the 3 variables (MII number, prescribed dose, and recommended dose), first, values for key combinations were set and reviewed with expert clinicians. Specifically, for 5 specific values of MII (0, 6, 10, 15 and 25), a table was crafted with the scores (from -1 to 1) describing the effect of, given a real dose rank of FSH and its outcome, changing it for another rank (or maintaining it). For example, if the outcome was 0 MII and the dose prescribed 100-150 IU/ml, maintaining that dose as a recommendation would result in a  $\varphi$  of -1, as the dose is clearly insufficient. Increasing the dose by one dose rank would get the recommendation to a  $\varphi$  value of -0.2, or in other words, closer to 0. The dose rank with the  $\varphi$  closer to 0 in this example would be 225-250 IU, as the case calls for a big change, but caution is introduced by giving a slightly worse value to the highest dose rank. All these values can be seen in Table 3.2 and Figure 3.2.

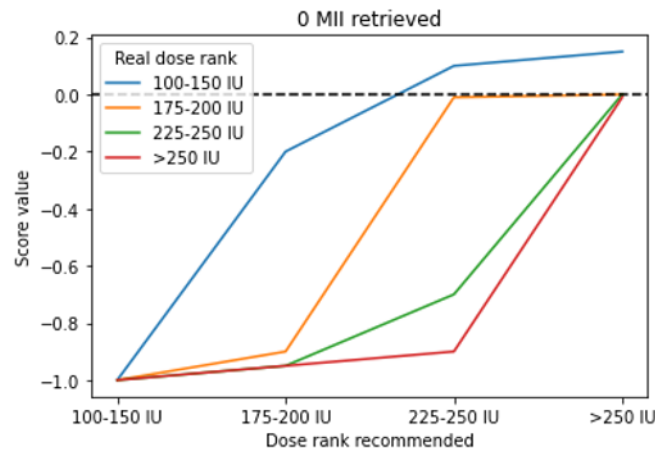


Figure 3.2: Linear representation of  $\varphi$  for all combinations of prescription/recommended dose ranks given that the outcome was 0 MII retrieved.

If the outcome  $y_i$  of a given case falls outside of the 5 key values,  $\varphi$  is computed by assuming a linear function between the available lower and higher key values that have an assigned  $\varphi$  value. This can be expressed as

Recommended dose rank	Real dose rank			
	100-150 IU	175-200 IU	225-250 IU	>250 IU
100-150 IU	-1	-1	-1	-1
175-200 IU	-0.2	-0.90	-0.95	-0.95
225-250 IU	0.1	-0.01	-0.7	-0.9
>250 IU	0.15	0	0	-0.01

Table 3.2:  $\varphi$  values for every prescribed/recommended dose rank given that the result was 0 MII

$$\varphi_i(y_i = a, \hat{d}_i, d_i) = \varphi(y_i = a^-, \hat{d}_i, d_i) + \frac{(\varphi(y_i = a^+, \hat{d}_i, d_i) - \varphi(y_i = a^-, \hat{d}_i, d_i)) * (a - a^-)}{(a^+ - a^-)}. \quad (3.2)$$

Where  $a^-$  is the key value of  $y$  immediately under  $a$ , and  $a^+$  the key value of  $y$  immediately over  $a$ . For example, if a case  $p_i$  has a value for  $y_i$  of 3, or  $a = 3$ , which is between the key values 0 ( $a^-$ ) and 6 ( $a^+$ ), receives a prescribed dose between 100 and 150 IU and it's recommended the same dose rank,  $\varphi$  would be computed as

$$\varphi_i(y_i = 3, \hat{d}_i = 100 - 150, d_i = 100 - 150) = -1 + \frac{(-0.8 - (-1)) * (3 - 0)}{(6 - 3)} = -0.9. \quad (3.3)$$

The equation is solved using  $\varphi$  values for those prescribed-recommended doses available in the tables for  $a^- = 0$  (Table 3.2) and  $a^+ = 6$  (Table 3.3). Tables A.3 to A.5 contain the rest of  $\varphi$  values for the key combinations that were used to construct the performance score ( $\varphi$ ) function. The values can be analyzed visually in Figures A.2 to A.5. These are available at Appendix A.

Recommended dose rank	Real dose rank			
	100-150 IU	175-200 IU	225-250 IU	>250 IU
100-150 IU	-0.8	-0.9	-0.95	-0.99
175-200 IU	-0.05	-0.60	-0.85	-0.9
225-250 IU	0.3	0	-0.5	-0.85
>250 IU	0.4	0.1	0	-0.001

Table 3.3:  $\varphi$  values for every prescribed/recommended dose rank given that the result was 6 MII.

Finally, in order to evaluate collective performance in comparison to clinical standard practice, the main interest is to assess which set of doses achieve a mean absolute  $\varphi$  closer to 0. This collective performance score can be expressed as

$$\Phi = \frac{\left| \sum_{i=1}^N \varphi_i(y_i, y^*, \hat{d}_i, d_i) \right|}{N}. \quad (3.4)$$

### 3.2.2 Prediction model

An ML model was trained to predict  $\varphi$  for all four dose ranks, given a specific patient. The variables age, BMI, AFC, AMH and proven fertility (yes/no) were used as predictors. For any case present in our database 4  $\varphi$  values were possible (one for each possible dose rank, the one prescribed and 3 possible recommendations). Thus, a data augmentation technique was passed on our database, resulting in its quadruplication, with every case (described always by the same predictor variable values) related to every 4 possible dose ranks and its corresponding scores ( $\varphi$ ). The performance scores here, are in fact evaluating counterfactual dose scenarios for each case. There is not sufficient information in the database to achieve precise counterfactual dose-response evaluation (as if there was enough information, causal inference methods could handle the problem), but clinical knowledge can complement it enough as to assess broadly whether a dose change is beneficial or detrimental.

These  $\varphi$  values were then considered to be the objective variable to be learnt by the algorithm. Before training the model, the variables containing the real stimulation dose rank and the number of MII recovered were dropped, as to hide them from the algorithm, making it only learn from the profile of the patient, the dose rank assigned, and the  $\varphi$  value calculated by our function.

A random 80% of the quadrupled database (conserving always the same patient together) was selected for training. The remaining 20% consisted of only the original cases without data augmentation, and was reserved for validation of the dosing model. The algorithm selected to learn this regression problem was a Random Forests Regressor (RFR). 5-fold cross-validation was performed.

### 3.2.3 Dosing model and performance evaluation

The resulting dosing application used the trained model to predict  $\varphi$  values for all 4 dose ranks for new patients, selecting the dose rank where the predicted  $\varphi$  value was closer to 0. This meant that, for each patient, the model predicted the best possible result (or  $\varphi$  closer to 0) with a specific dose rank. Thus, that dose rank was the one recommended by the dosing model. This process was executed for all 5 randomly selected test datasets during the development phase, and again for the cases in the time separated validation dataset.

To evaluate the performance of the dosing model compared to clinical practice both in test and validation, the score function was used to compute the  $\varphi$  values of the recommendations outputted by the model, as the real prescribed dose ranks and the outcome in number of MII in those datasets were also available.  $\varphi$  and  $\Phi$  for the recommendations were compared to those computed for clinical prescriptions, graphically in the first case and statistically in the second. For the statistical comparison, mean absolute values, or  $\Phi$ , were compared using Wilcoxon signed-rank test, as the

distribution of the  $\varphi$  was not normal and the values compared were paired (related to the same set of cases). The hypothesis 1 was that of a one-sided difference between means, specifically, that the value of  $\Phi$  for the recommendations by the dosing model was smaller, or in other words, closer to 0.

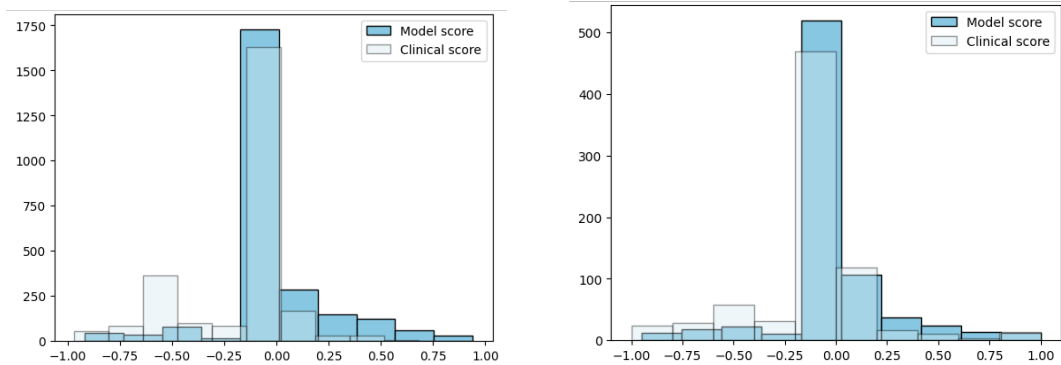
### 3.2.4 Results

The mean absolute scores ( $\Phi$ ) for the dose rank recommendations of the model were significantly lower both in development and validation phases (Table 3.4).

	Clinical prescription	Model recommendation	p-value
$\Phi_{development}$	0.17 (95% CI 0.16 to 0.18)	0.12 (95% CI 0.11 to 0.12)	<0.01*
$\Phi_{validation}$	0.16 (95% CI 0.14 to 0.18)	0.13 (95% CI 0.11 to 0.15)	<0.01*

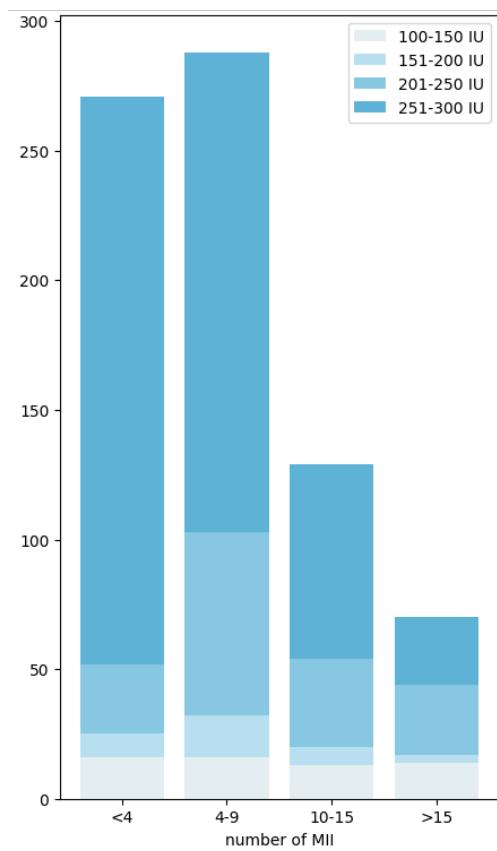
Table 3.4: Mean absolute score ( $\Phi$ ) values plus 95% confidence interval (CI) for clinical dose rank prescriptions and model recommendations during development and validation phases. Statistical differences tested using the Wilcoxon signed-rank test. A p-value under 0.05 was considered significant.

Distributions of the signed scores ( $\varphi$ ) during development and validation can also be visualized in Figures 3.3a and 3.3b. In both cases a higher accumulation of  $\varphi$  values around 0 is observed for the model recommendations versus the clinical prescriptions. This distribution agrees with the statistical difference observed.

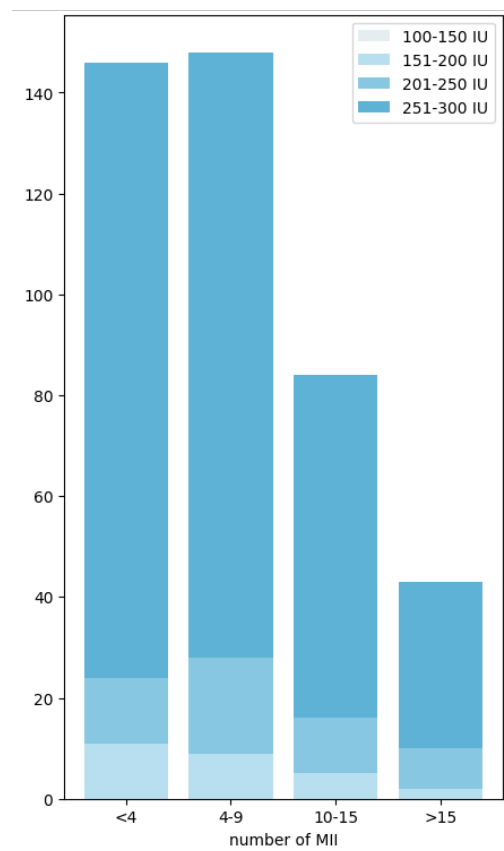


(a) Distribution of scores for clinical prescriptions ( $\varphi_c$ ) and model recommendations ( $\varphi_m$ ) in the development database. (b) Distribution of scores for clinical prescriptions ( $\varphi_c$ ) and model recommendations ( $\varphi_m$ ) in the validation database.

Further, when dose ranks assigned by the model are compared to those by clinicians per range of MII obtained, sub-optimal and poor response patients are assigned a clear increment in dose rank (Figures 3.4a and 3.4b). On the other hand, an increment is also observed in patients with optimal and hyper-responses. This does not negate the collective improvement, especially in poor and sub-optimal response cases, but it is not to be disregarded.



(a) Dose ranks prescribed per range of MII retrieved by the clinicians during validation.



(b) Dose ranks prescribed per range of MII retrieved by the model during validation.

### 3.2.5 Lessons learned

Upon construction of the initial iteration of the FSH dosing model and subsequent analysis of its performance through the developed score, it was apparent that the two challenges introduced at the beginning of the chapter could be overcome. First, the model was capable of recommending higher doses for expected poor responders ( $\text{AMH} < 1.2 \text{ ng/ml}$  and  $\text{AFC} < 5$ , as per POSEIDON criteria Esteves, Roque, et al., 2018), and lower for expected high responders ( $\text{AMH} \geq 3 \text{ ng/ml}$  and  $\text{AFC} \geq 15$ ), even if with some exceptions. Specifically, it recommended expected poor responders with the highest dose rank in a 100% of cases, and expected high responders with the remaining lower rank doses in a 84.6% of cases.

Therefore, the developed score facilitated the circumvention of the confounding effect present in the dataset by indirectly training the model to recognize the positive dose-response relationship. This is reflected in the assigned score, as for example, for cases with poor responses ( $< 4 \text{ MII}$ ) a higher dose rank gets a  $\varphi$  closer to 0 (the more dose, the higher effect, which is needed in this case). Of course, the score function also included penalization ( $\varphi$  not as close to 0) whenever it was deemed that although a change in dose was necessary, the dose rank recommended was too far removed from the clinical dose rank. This was done to introduce a measure of precaution, because when the outcome is already close to the optimal range, large changes are more likely to be detrimental.

In conclusion, given that informing indirectly the ML model about the nature of the studied dose-response relationship allowed for a clinically consistent model, it was expected that specifying this relationship more clearly could improve the performance. Furthermore, following this strategy would improve interpretability of the resulting model. In the next section a second iteration is described, where the predictive model is based in an assumed linear dose-response relationship.

## 3.3 Second iteration

As explained in earlier chapters (Chapter 2, subsection 2.2.2), available research suggests that a sigmoid function could explain closely the dose-response relationship between the dose of FSH administered and clinical outcomes like number of oocytes retrieved after OPU. Ideally then, finding the 4 parameters of a Hill's equation (or the 3 of a logistic curve) for each individual patient would take us closer to a model based in the pharmacometric characteristics of FSH. In other words, closer to physiological reality. But to be able to approximate these curves for every patient, multiple dose-response points from each individual case are needed. Or at least, diverse dose prescriptions for similar patients with its outcomes. The first requirement clearly is difficult to obtain, as the majority of patients get only one or two treatments, as they either achieve their objective (a baby at home), do not have the economical or psychological means to endure more treatments,



or decide to change the type of treatment. The second requirement poses a significant challenge due to the clinicians' adherence to standard clinical practice, as previously noted. Therefore, the available dataset did not lend itself to the optimal approximation of individualized sigmoid curves.

Sigmoid curves have two main characteristics of interest in this context:

- *Saturation*: Property that describes the asymptotic behavior of the curve as its input approaches positive or negative infinity.
- *Monotonicity*: Property that describes the direction and consistency of a function's output as its input changes. In the context of positive monotonicity, the function's output is non-decreasing as the input increases, meaning that it either increases or remains constant, but never decreases.

A linear function complies with the monotonicity property, and it is composed by only two parameters that need to be estimated: the intercept and the slope. In the next sections a thorough description of how individualized linear dose-response functions are estimated using observational databases with only one dose-response data point per patient can be found.

### **Predictive model construction**

A linear function can be mathematically described as:

$$y_i = y_0 + s_i * d_i \quad (3.5)$$

Where  $y_i$  is the outcome (in this case the number of MII) of a patient  $p_i$ ,  $y_0$  the intercept or outcome whenever input is 0,  $s_i$  the slope of the line, and  $d_i$  the input (here the dose of FSH).

For any patient, during a natural cycle (0 IU/ml of exogenous FSH) the outcome in number of MII collected would stay mainly between 0 and 1. This would describe the intercept parameter of a linear dose-response function. Given that the value range of the intercept hardly varies from patient to patient, it could be assumed equally for all population. Hence, the slope parameter of the function, which describes the patient's capability of reacting to the first dose of FSH, was the only one that needed to be computed individually. As the results of a specific dose of FSH for each case were available in the databases, the value of individual slopes was easily calculated. To avoid negative slope values, it was assumed that all patients would achieve 0 MII if given 0 exogenous FSH, thus the intercept was fixed to 0.

The slope of a linear function is defined as follows:

$$s_i = \frac{y_i - y_0}{d_i - d_0} = \frac{y_i}{d_i} \quad (3.6)$$

As the first data point  $(x_0, d_0)$  can be set at the origin  $(0, 0)$  due to the intercept being fixed at 0, the slope value for every patient is computed by dividing the outcome or  $y_i$  (MII) by  $d_i$  (the first dose of FSH). Thus, the slope was calculated for each case in the development database.

The slope value calculated was set as the target variable to be learnt, and a linear regression algorithm was trained to predict it for every case (defined by its values at the start of the stimulation in age, BMI, AFC, AMH and proven fertility). Training was conducted on a random 80% of the development database. The remaining 20% was reserved for testing purposes. The training process was again cross-validated five times with five randomly selected training datasets, with their corresponding five test sets.

### **Dose recommendation by the model**

For dose-recommending purposes, the predicted slope for each test patient was used to compute the necessary FSH to obtain an outcome of 12 MII (middle point for the 10 to 15 optimal range) using the dosing function, derived from the linear function:

$$\hat{d}_i = \frac{y^*}{\beta^T x_i} \quad (3.7)$$

Where  $\hat{d}_i$  is the FSH dose recommended for the case described by  $x_i$  variables,  $y^*$  the desired outcome, and  $\beta$  the set of coefficients estimated by the linear regression model trained to predict the slope using the covariates included in  $x_i$ . Thus,  $\beta^T x_i$  represents the value of the predicted slope. Again, recommended doses were capped at 300 IU/ml.

### **Evaluating the performance of the model**

The performance of model-based recommendations was evaluated using the  $\varphi$  function in the 20% portion reserved for testing in the development database (cross-validated 5 times) and in all the cases of the validation database. In both cases, two  $\varphi$  values were computed for each patient. One for the dose prescribed by the clinician ( $\varphi_c$ ) and another for the model recommended dose ( $\varphi_m$ ). Mean absolute values of both set of scores ( $\Phi_c$  and  $\Phi_m$ ) were compared across all cases to identify which group (clinical or model recommended) had  $\Phi$  closer to 0, being of no importance if the dose was too high or too low, just as in the previous iteration. Again, the Wilcoxon signed-rank test was used for this purpose.

### 3.3.1 Results

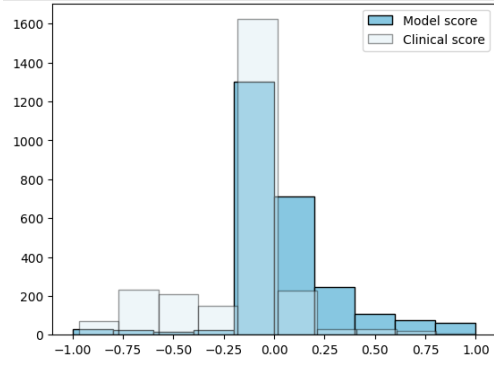
#### Predictive and recommendation performance

Both in development and in validation phases, the dosing model created had a significantly lower  $\Phi$  compared to dose ranks prescribed by clinicians (see Table 3.5). When comparing the results from this iteration ( $\Phi_2$ ) to those of the initial iteration outlined in preceding sections ( $\Phi_1$ ), the findings indicated that there was no statistically significant difference in the results attained from the development database (p-value = 0.098). However, a significant enhancement from the second iteration as compared to the first was observed in the validation database (p-value < 0.05).

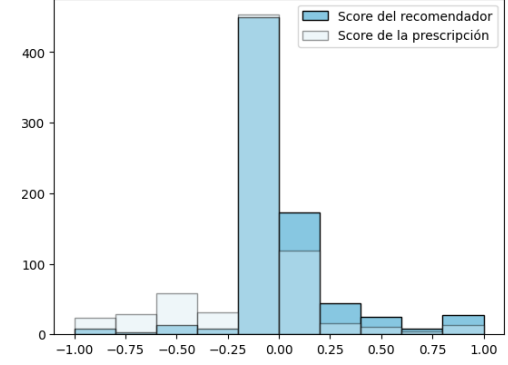
	Clinical prescription	Model recommendation	p-value
$\Phi_{development}$	0.17 (95% CI 0.16 to 0.18)	0.13 (95% CI 0.12 to 0.14)	<0.01*
$\Phi_{validation}$	0.16 (95% CI 0.14 to 0.18)	0.11 (95% CI 0.10 to 0.12)	<0.01*

Table 3.5: Mean absolute  $\varphi$  values ( $\Phi$ ) plus 95% confidence interval (CI) for clinical dose rank prescriptions and model recommendations during development and validation phases. Statistical differences tested using the Wilcoxon signed-rank test. A p-value under 0.05 was considered significant.

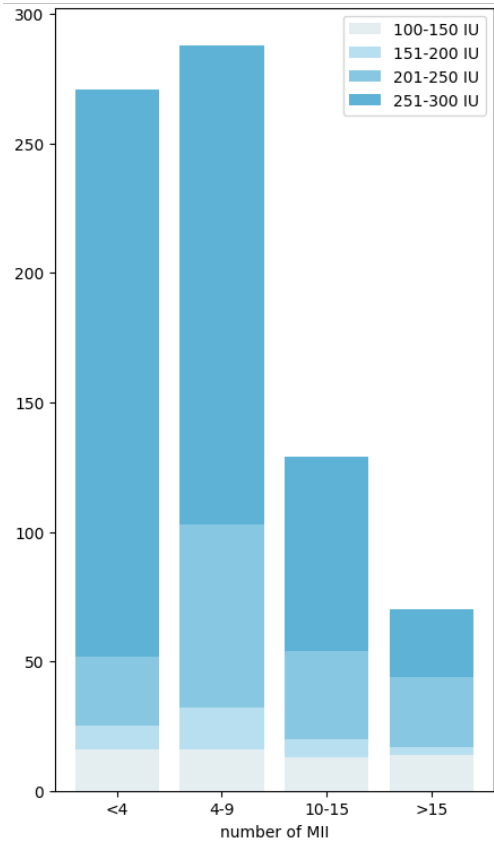
To further understand the performance of the model and of the clinical prescriptions, the model performance was analyzed graphically as in previous sections. The clinicians' score distribution ( $\varphi_c$ ) were compared to that of the scores from the model ( $\varphi_m$ ) in the test sets of the development database and in the validation one (Figures 3.5a and 3.5b). The model's  $\varphi$  values approached 0 (the best possible dose rank) more times than the clinicians' scores, suggesting dose ranks higher than the one favored by clinicians when their  $\varphi$  values were not approaching 0. In 57.4% of cases in the test set and in 68.8% in the validation database, the dose rank was not modified in relation to the clinician-prescribed dose. How the dosage was changed from clinician prescription to model recommendation was further analyzed in relation to the real outcome in number of MII in Figures 3.6a and 3.6b. Visual analysis shows that the model tends to increase the dose for patients with low and sub-optimal oocyte retrievals, but also increases dosage for some of the hyper-responders. Tendencies are similar between this model and the first iteration, with a 100% of expected poor responders receiving a recommendation of the higher dose rank, and a 89.7% of expected high responders receiving the remaining lower dose ranks. This means that the current model recognizes slightly better the expected high-responder profile, even if not as accurately as expected poor responders. This leads to the already mentioned increase in dose rank in some hyper-responders (>15 MII obtained).



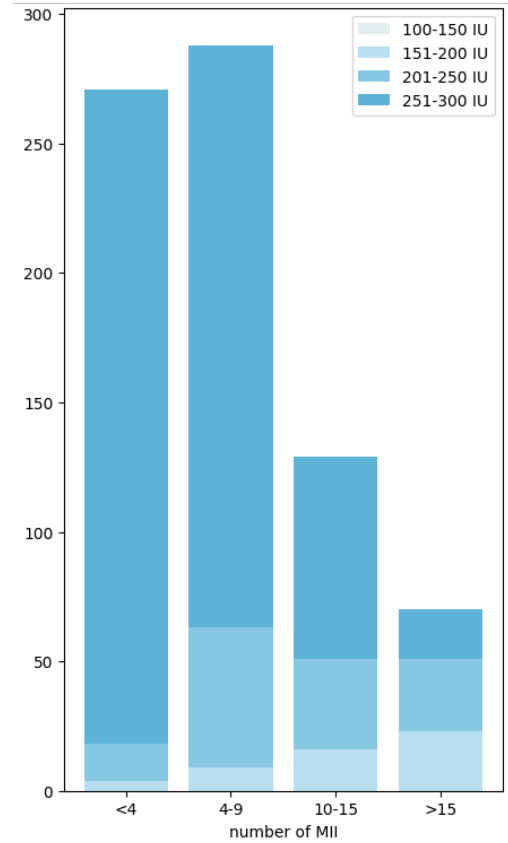
(a) Distribution of scores for clinical prescriptions ( $\varphi_c$ ) and model recommendations ( $\varphi_m$ ) in the development database.



(b) Distribution of scores for clinical prescriptions ( $\varphi_c$ ) and model recommendations ( $\varphi_m$ ) in the validation database.



(a) Dose ranks prescribed per range of MII retrieved by the clinicians during validation.



(b) Dose ranks prescribed per range of MII retrieved by the model during validation.

### 3.4 Discussion

Currently, literature on FSH dosage models includes several recommending models that have provided promising results. Yet, some of them have not been tested by RCT; those that have (Olivennes et al., 2015; Allegra et al., 2017; Nyboe Andersen et al., 2017), however, have not been developed for use on all types of patients. The inclusion of only normo-ovulatory patients (Howles et al., 2006), or patients younger than 40 years with regular cycles (La Marca et al., 2012; Nyboe Andersen et al., 2017) restricts the personalization of the first FSH dose to a small subset of patients. In this subset, however, this personalized dose finding is not as critical as for the excluded patients. As the models presented in this chapter include every type of patient, the results are enhanced for all of them.

In addition to age, which is the strongest predictor; AFC, AMH, BMI and presence of previous successful pregnancies as variables in the core model have been shown to be good predictors of the dose-response function slope. This slope value has already been used as ovarian sensitivity (oocytes recovered per unit of starting FSH) in the development of a monogram tested by RCT (La Marca et al., 2012). Its use as an objective variable of the core model mitigates the confounding effect produced in any non-randomized treatment database, and that could lead a direct model (oocyte number as objective variable) to determine, for example, that higher doses, which are often prescribed for low-responders, lead to smaller oocyte yields. As already mentioned, the treatment is tailored to the patient by the clinician, and it is especially important to account for the confounding it can cause in a non-randomized database. Removal of this effect also leads to a model adherent to clinical knowledge, understanding that extreme patients (low-responders and hyper-responders) have extreme ovarian potential values.

During this chapter we have reviewed two incremental iterations of FSH dosing models, along with the construction of a score function able to evaluate dose recommendations pre-clinically, tackling successfully the two challenges set at the beginning of the section. The second iteration, based on an assumed positive linear dose-response relationship, demonstrated slightly better performance than the first. Thus, up until this chapter, it is the preferred version. Not only did it obtain better  $\varphi$  values during the validation phase, its construction around a linear function allows for an easier interpretation of how the model is assigning doses for each patient. In Figure 3.7 three linear dose-response functions are plotted, each with a different slope value, and with markers indicating the dose needed to achieve  $y^*$  (12.5) in each case. With the slopes values of 0.15 and 0.07, the doses recommended by the function are inside the allowed space, but with the lowest exemplified slope value, the maximum allowed dose (300) does not reach  $y^*$ . This is not only easy to understand, but does reflect also faithfully the physiological reality of poor responder patients.

Further, the continuous prediction of the doses, as opposite to dose ranks by the first iteration, allows for an even finer tailoring of the FSH dose. Of course, full continuous selection of FSH dose

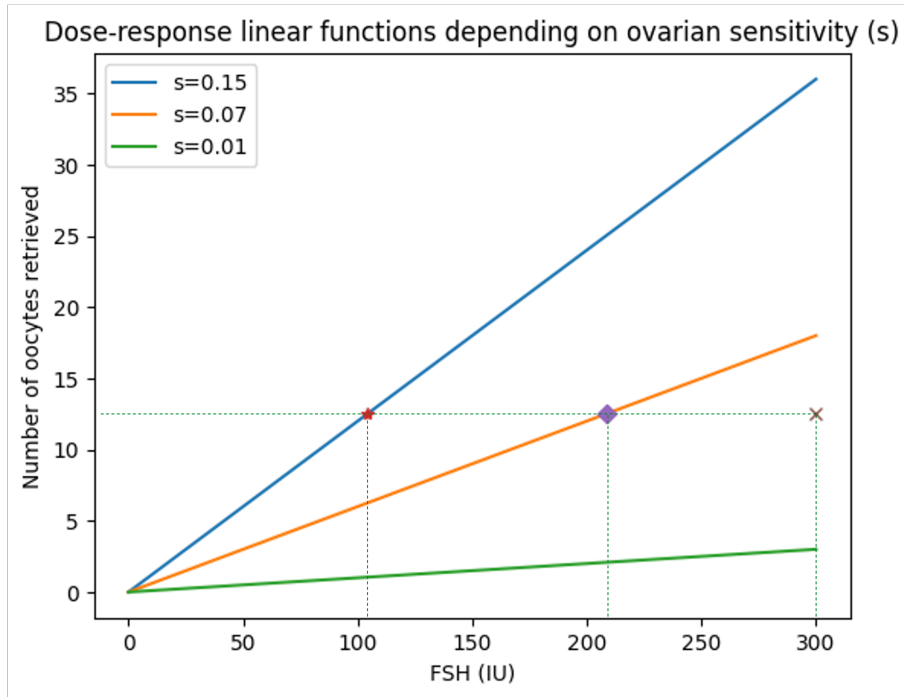


Figure 3.7: Graphical representation of three individualized linear dose-response functions with three different slopes. The marked points indicated the dose needed to achieve 12.5 mature oocytes for each function.

for injection is not realistic, as there are some limitations to the granularity of drug administration, but real dose steps are much finer than the dose ranks proposed in the first iteration. Hence, a model that outputs continuous recommendations can be easily adapted to the real fine dose-steps. Additionally, the dosing function enables the final user to select the number of MII desired to be retrieved, and then obtain the corresponding FSH dose recommendation. This opens the use of this model to all kinds of situations, not just those in which 12 MII, or 10 to 15, are the desired result.

As a separate contribution to an inclusive FSH dosing model, the developed performance score allows to test in silico whether the model would improve results compared with historical data, as a step preceding an RCT. To this end, the performance score was designed to encode and automate faithfully an expert clinical assessment of treatment-recommendation-outcome combinations. In other words, it allows to test whether a recommended dose could fare better than the one already prescribed, given the real result in retrieved MII. It is therefore possible to estimate reliably whether the model can improve current clinical practice. With this information, the investment in a well-designed RCT can be made more confidently. Additionally, results of the in-silico performance of the model are more informative than the sole prediction scores (like R squared) of the core dosing model.

The  $\varphi$  values of the selected model were consistently better than those of clinical practice, both

in the development and validation databases. This is of interest as the model holds its value even though the population of the validation database is significantly older. Therefore, it means that the core model has learnt the important aspects of the relationship between the patient's characteristics and her ovarian potential or slope in the dose-response function. It is worth noting that the most significant predicted improvement was for those patients whose oocyte yield was low or sub-optimal, in which doses are increased on average. Upon implementation, the system's recommendations may improve the average results and most probably avoid some cycle cancellations owing to lack of embryos for transfer.

Detailed analysis of the behaviour of both iterations revealed their tendency, when incorrect, to overdose some patients. This contrasts with clinical practice, in which the tendency is to underdose when the prescription is inadequate. These instances of overestimations by the models correspond mainly to hyper-responder patient profiles, which are under-represented in our databases. As such, the algorithm could not learn appropriately owing to the lack of a sufficient sample size. Importantly, although the selected model does tend to overdose these patients, it still recommends the same or lower doses than the clinician in most of these cases, i.e. the clinician also tends to overdose. Nonetheless, we cannot dismiss the possibility that this could lead to a small increase in the risk of OHSS. This contrasts with previously published results in which RCT-tested models reduced the incidence of OHSS risk (Olivennes et al., 2015; Allegra et al., 2017; Nyboe Andersen et al., 2017). Secondary results of these studies, however, failed to show an increase in either retrieved oocytes or pregnancy results, with one reporting a reduction in oocyte yield (Olivennes et al., 2015). Although the risk of OHSS must be taken seriously, it is also true that it can be managed within a cycle with proper prevention, such as gonadotrophin releasing hormone agonist trigger and deferred embryo transfer. All things considered, it is not unreasonable to consider a manageable, or even almost completely avoidable, risk for a small portion of patients in order to avoid a lack of embryos suitable for transfer for others.

Further analysis of the instances in which the selected model made a sub-optimal suggestion led to another conclusion. Instances in which the model had negative  $\varphi$  values seem to coincide frequently with negative values for the clinician's prescription. Analysis of these cases in more detail produced a profile of patients with good markers and an unexplained low retrieval of oocytes. This could possibly be related to undiagnosed genetic polymorphisms in the FSHR or LHB genes (Lledo et al., 2014), which, obviously, neither the clinicians nor the model could detect. Despite the possible limitations of the system, it is encouraging that the preliminary results show, in most cases, similar or better  $\varphi$  of the model's recommendation compared with the dose prescribed by the clinician.

In conclusion, clinicians prescribe the first FSH dose for each patient based on their characteristics, reserve markers and their own experience with similar cases. Although most of the time they prescribe the dose necessary for an optimal result, sometimes the outcome can unexpectedly vary

and fall into sub-optimal or extreme ranges. Our model could avoid most of these deviations by analyzing the patient's profile and making suggestions for the medical professional to assess. Once tested and its performance confirmed by RCT, the ML selected model could be used as a training and learning tool for new clinicians and could serve as quality control for experienced ones; furthermore, it could provide a second opinion as the information could be useful in peer-to-peer case discussions.

### 3.5 Conclusions

At the beginning of this chapter we set up to answer the first research question of this project:

**Q1:** *"Is it possible to improve clinical FSH dosing policy for Controlled Ovarian Hyperstimulation using only historical data?"*,

which needed first an answer for its subquestion

**Q1a:** *"How can we analyze a dosing model's performance before clinical intervention?"*.

The first research question (**Q1**) has been given a positive answer, with a dosing model based in an assumed linear dose-response function. As for its subquestion **Q1a**, the answer passed through the injection of coarse grained counterfactual evaluations constructed as a tailored performance score to evaluate all possible FSH doses for each case. This demonstrates that the shortcomings of observational datasets (low treatment allocation variability and difficulties to have complete information on all relevant variables) can be surpassed if field knowledge is included in the construction of the dosing models. Of course, more balanced and varied datasets would enhance results (for example with more datapoints for hyper-responders), as would the introduction of additional and more precise clinical knowledge.

The performance score has played a crucial role in evaluating the potential improvement of the dosing model versus a clinical reality, but it has been developed as an ad hoc solution. This means that it is not extendable to other similar dosing problems in its present form. In order enable its extension to other dosing problems, a general method needs to be devised. In the next chapter, a generalizable solution is presented and demonstrated in the same FSH dosing problem.



## Chapter 4

# IDoser: including field knowledge into the training of dosing models

In this chapter, we will present an approach to answer the second research question

**Q2:** *"Can we extend this methodology to other dosing problems?"*,

by presenting a novel methodology and applying it to the FSH dosing case.

The chapter will cover:

- A formal definition of a general one-time dosing problem
- A thorough description of the proposed methodology
- An implementation in the FSH dosing problem with positive results
- Lessons learned and conclusions

This chapter is a lengthier description of the article "Núria Correa, Jesús Cerquides, Rita Vassena, et al. (2023). "IDoser: Improving individualized dosing policies with clinical practice and machine learning". In: *medRxiv*. DOI: 10.1101/2023.03.28.23287859" which has also been accepted as an oral communication in the ESHRE (European Society for Human Reproduction and Embryology) annual meeting of 2023 in Copenhagen as "Núria Correa, Jesús Cerquides, Josep Lluís Arcos, Rita Vassena, and Mina Popovic (2023a). "O-185 A clinically robust machine learning model for selecting the first FSH dose during controlled ovarian hyperstimulation: incorporating clinical knowledge to the learning process." In: Oral communication. European Society of Human Reproduction and Embryology (ESHRE) Annual Meeting. Copenhagen".

## 4.1 Background

As previously explored in Chapter 2, there are methods available to optimize a drug dose selection. Either PX methodologies or causal inference, separately or combined, especially when enhanced by ML, are appropriate approaches that can provide model informed precision dosing (MIPD). The advantages and disadvantages of said methods are briefly summarized here: on the one hand, PX's mathematical methods rely heavily on known physiological properties of dose-response relationships, which delivers clinically trustworthy models. These models are obtained and further refined using prospective data. The utilization of prospective data obtained from randomized trials enables a more thorough investigation of dose-response functions, as the data are expected to be independent from confounding factors. However, for social, practical and financial reasons alike, trials have been historically often limited to an specific portion of population and lack enough diversity as to have its results be applicable to a diverse population (Oh et al., 2015; Keizer et al., 2018). Further, trials often exclude patients with certain comorbidities (sometimes all of them) that are relevant for the studied drug-response function (Gonzalez et al., 2017). Lastly, it is also not uncommon to find PK/PD models fitted for a proxy outcome rather than the clinically relevant one due to time and cost considerations. To summarize, frequently PX models available for specific drugs are not suitable for clinical practice, or are only fit for certain portions of the patient population. Methodologies to adapt these models for a more diverse population and considering relevant biomarkers require either prospective data, an initial model fitted to a clinical relevant outcome and related to at least some biomarkers, or access to data on drug concentration in blood plus a diverse and complete observational database (in order to develop a new model without prospective data). For the reasons explained above and in Chapter 2, subsection 2.1.1, these requirements are overall difficult to comply with.

On the other hand, causal inference methods do not entail dependence on a PX model fitted to relevant biomarkers and/or clinical outcome, while they do rely on the causal interdependence of covariates, drug and outcome. By design, it enables the use of confounded observational databases in order to obtain reliable causal models. Hence, it maintains causal sense, close to physiology, and does not need prospective data. However, as mentioned in previous chapters, there are two strong assumptions that need to be satisfied for these methods to work properly: unconfoundedness and overlap. In practice this means that (1) all confounders are accounted for and (2) that all cases have a non-zero probability to receive all treatments or doses possible. For the first statement, it is important to note that a common struggle in EHR databases is that of completeness. As clinicians will have diverse medical criteria for which biomarkers are relevant, not all patients will get tested for all known confounders. Furthermore, the economic costs associated with testing may impede the fulfillment of all necessary tests. For the second, clinical adherence to standard protocols is typically high, indicating that patients who fall under certain categories are unlikely to receive treatments or doses that are not recommended for them. Hence, day-to-day clinical practice rarely

produces observational datasets that are amenable to the use of current causal methodologies.

This leaves a gap for situations where drug dosing policies are sub-optimal and would still need an improvement. In Chapter 3 we already explored an ad hoc solution for the FSH dosing policy, that partly takes inspiration in both PX and causal methodologies by introducing a core model that complies with known physiological and/or causal characteristics. In the next sections a general extension of this idea, along with a generalization of the dose evaluating system is presented and tested in the FSH dosing policy optimization problem. The common objective to reach a 10 to 15 oocyte range stays the same, but maximal dose is explored up to 450 UI.

## 4.2 The Individualized Dose Improvement Problem

The problem under study can be summarized as the Individualized Dosage Improvement Problem (IDIP). Given a large population of  $N$  patients  $P$ , the goal of an IDIP is to select the optimal quantity of a certain drug, which we refer to as the *dose*. For every patient  $p_i \in P$  a *dose* and its outcome or *response* is recorded, and there is only a pair of *dose-response* values. We represent the response of  $p_i$  by a real number which we refer to as  $y_i \in \mathbb{R}$ , and the dose as  $d_i \in [0, \infty)$ . We assume that the response can be measured by a single real number. Furthermore, we assume that the desired levels of response are known for each individual, which we describe as  $y_i^*$ .

Each patient  $p_i \in P$  is described by a set of  $k$  characteristics  $x_i = (x_i^1, \dots, x_i^k) \in \mathcal{X} = \mathbb{R}^k$ . These characteristics can include for example, the patient age in years, its weight, height, gender, values of previous analysis, and so on.

An *individualized dosage policy*, or IDP,  $\pi : \mathcal{X} \rightarrow \mathbb{R}$  is a function that decides a *proposed dose*  $\hat{d}_i$  provided the characteristics of the patient ( $x_i$ ).

The objective is to find an IDP or  $\pi$  with the minimum error or *loss* possible. Mathematically, this means identifying

$$\pi^* = \arg \min_{\pi} L(\pi) \quad (4.1)$$

where  $L$  is a *collective loss* function. That is, the higher the  $L(\pi)$  the smaller the quality of the IDP  $\pi$ . The collective loss is computed as the average of *individual losses* ( $l_i$ ), one for each patient  $p_i$ . The individual loss of a dose on a patient  $l(y_i, y_i^*, \hat{d}_i, d_i)$  measures how good a proposed dose ( $\hat{d}_i$ ) would be depending on the real  $d_i$ , its correspondent  $y_i$  obtained and the objective  $y_i^*$ . Thus, the *Collective loss* can be mathematically defined as

$$L(\pi) = \frac{\sum_{i=1}^N l(y_i, y_i^*, \pi(x_i), d_i)}{N}. \quad (4.2)$$

In the IDIP, we are provided with data that describes current practice for the dose policy of the

drug. Specifically, we are provided with information about  $N$  patients out of the complete population, and for each patient  $p_i$ ,  $i \in [1..N]$ , that has been administered the drug, we record

- their characteristics  $x_i \in \mathcal{X}$ ;
- the dose of drug administered to this patient, namely  $d_i \in [0, \infty)$ ;
- the response value obtained, namely  $y_i \in \mathbb{R}$ .

The main challenge posed is: How can we use the information available from current dosing practices to create an individualized dosing policy with minimal loss?

### 4.3 Proposal: Individualized Doser (IDoser)

Our proposal hinges on the following two assumptions:

1. The dose-response function is at least locally *monotonic*, that is the larger the dose, the bigger (or equal) the expected response.
2. There is a known *optimal outcome* ( $y^*$ ) and it is known to us. This can either be a range or a point.

*IDoser* is then constructed around two elements: (1) A **Core dosing model** that relates  $X$  to  $d$  through a set of coefficients that we describe as  $\gamma$ , and is used to predict  $\hat{d}$ ; and (2) A **loss function** that evaluates the predicted  $\hat{d}$  depending on  $d$  and  $y$  (see Figure 4.1).

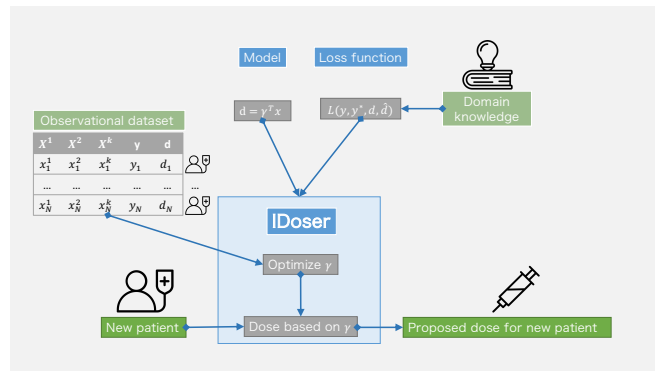


Figure 4.1: Principal components of IDoser.

We will review both elements in the following subsections. For the rest of the manuscript positive monotonicity will be the default assumption. Nevertheless, IDoser can be effortlessly adapted for a negative monotonicity assumption.

#### 4.3.1 The core model

Given that our main interest lies in predicting the optimal dose for each patient, a general and parametric core model is represented as follows:

$$\hat{d}_i = \pi_\gamma(x_i) \quad (4.3)$$

Where  $\gamma \in \mathbb{R}^\kappa$  is the parameter.

A core model can be specified in linear form as:

$$\hat{d}_i = \pi_\gamma(x_i) = \gamma^T x_i \quad (4.4)$$

This is the simplest form that complies with the monotonicity assumption and the requirement of relating the dose  $\hat{d}_i$  to the covariates  $x_i$  through  $\gamma$ , but other more complex forms can be used as needed.

#### 4.3.2 Loss function

Being able to evaluate a hypothetical or counterfactual dose is key for achieving an improvement on any real dosing policy ( $\pi$ ). Due to the limitations found in clinical observational datasets, we do not have enough information to build a predictive model for  $y$  in counterfactual doses, but we have enough expert knowledge to evaluate if the doses are good or bad based in ground truth found in the database. For this purpose, we codify field knowledge into the loss function. Namely, our two assumptions: positive monotonicity (general or local), and the existence of a desired outcome  $y^*$ . This can be easily introduced as follows:

$$l(y, y^*, \hat{d}, d) = \begin{cases} -1 & \text{dose change is correct} \\ +1 & \text{dose change is incorrect} \\ 0 & \text{no dose change} \end{cases} \quad (4.5)$$

where a correct dose change is increasing  $d$  whenever  $y < y^*$ , and decreasing  $d$  when  $y > y^*$ . Any change outside of these assumptions would be incorrect (Figure 4.2).

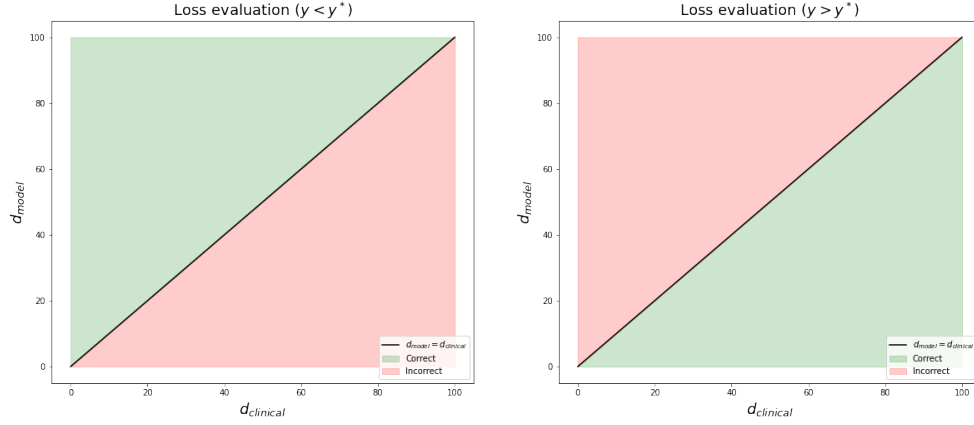


Figure 4.2: Graphical representation of loss evaluation for cases where  $y < y^*$  (left) and where  $y > y^*$  (right)

Given positive monotonicity, for any  $p_i$  that has  $y_i > y^*$ , an increase on dose would move  $y_i$  further from  $y^*$ , hence impairing the outcome. In the situation, an improvement would be to reduce dose. On the situation where  $y_i < y^*$  the reverse is true.

It is worth noting that the idea of local monotonicity enables inclusion of dosing settings where the general monotonicity is an extreme assumption. But if inside local monotonicity can still be assumed inside a specific dosing space, a local policy can be optimized using the method IDoser.

The function that generates the loss evaluation can be modified as needed for negative monotonicity and, further, complemented with additional rules that may be required depending on the situation or specific use cases.

One example of this is not considering any change in the right direction as good, and introducing limitations on dose change. This kind of limitations would ensure that uncertainty is considered, as larger changes in dose imply less confidence in its effect.

Another rule may involve introducing a certain threshold to start considering  $\hat{d}$  as different from  $d$ . These additional rules would consequently change our allocation of loss value as represented in Figure 4.3 and will be exploited in our use case.

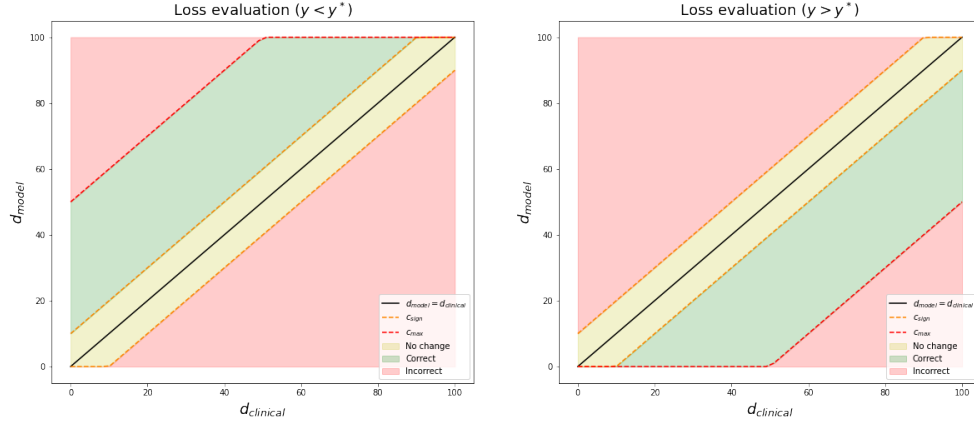


Figure 4.3: Graphical representation of loss evaluation for cases where  $y < y^*$  (left) and where  $y > y^*$  (right) with additional rules considering maximum change allowed and a minimum change threshold.

### 4.3.3 Optimization of parameters

After definition of the core model and the loss function according to the use case selected, the parameters of the model are found by minimization of the loss function. Here, we propose a coordinate descent algorithm (Wright, 2015), to iteratively establish the set of parameters that results in a minimum collective loss or  $L$ :

$$\gamma^* = \arg \min_{\gamma} L(\pi_{\gamma}) = \frac{\sum_{i=1}^N l(y_i, y^*, \pi_{\gamma}(x_i), d_i)}{N} \quad (4.6)$$

Once a minimum is reached within a randomly selected portion of the database (training), the resulting parameters  $\gamma^*$  determine the optimized dosing policy, namely  $\pi^* = \pi_{\gamma^*}$ .

## 4.4 Use case

The use case aims to find the right FSH dose in a COH for an IVF treatment. In the available observational dataset, a set of covariates related to the ovarian reserve of the patient are observed, together with the dose of FSH prescribed by clinicians and the outcome, measured in the number of mature oocytes retrieved. The covariates include: patient age at the time of treatment, body mass index (BMI), AFC, AMH levels, and basal endogenous FSH levels. Compared to the variables used in Chapter 3, basal FSH was added, and presence of previous fertility discarded. Presence or absence in the model of the variable previous fertility was assessed, concluding that its weight in the model was not very relevant. Regarding basal endogenous FSH levels, this variable was

	Development database (n = 7768)		Validation database (n = 273)	
<b>age</b>	$37.09 \pm 4.85$	[18-51]	$38.13 \pm 4.10$	[24-46]
<b>BMI</b>	$23.75 \pm 4.22$	[14.53-45.18]	$22.98 \pm 4.02$	[16.45-41]
<b>AFC</b>	$11.92 \pm 7.73$	[0-81]	$11.49 \pm 9.15$	[0-85]
<b>AMH</b>	$2.38 \pm 2.33$	[0.01-32.95]	$2.29 \pm 2.5$	[0.01-23.70]
<b>basal FSH</b>	$7.47 \pm 4.19$	[0.1-94.00]	$8.78 \pm 6.72$	[0.93-89.60]
<b>FSH dose</b>	$246.96 \pm 58.95$	[100-600]	$268.64 \pm 54.73$	[112.5-450]
<b>MII</b>	$7.30 \pm 5.26$	[0-47]	$6.55 \pm 6.07$	[0-36]

Table 4.1: Summary statistics of development and validation databases.

introduced in this experiment in order to analyze whether its inclusion in the model increased its performance compared to not including it.

Two databases were retrieved. One dedicated to developing the dosing models, composed of first IVF cycle patients undergoing treatment between January 2011 and December 2019; and a second one reserved only for validation of the resulting models (cases from January 2020 to September 2021). A summary of the characteristics of the two databases can be found in Table 4.1.

Thanks to available literature, we confirmed that both assumptions needed for our proposal hold true. For assumption 1 (positive monotonicity), while some evidences in cows may defy it (Karl et al., 2021), in the human species, no increment of FSH dose results in a lower number of oocytes retrieved under the same circumstances (same patient, same menstrual cycle) (Porchet, Le Cotonnec, and Loumaye, 1994; Arce, Klein, and Erichsen, 2016; Lensen et al., 2018; Abd-Elaziz et al., 2017). The only negative effects of higher doses of FSH observed in human relate to the quality of oocytes (Luo et al., 2022) and not their quantity. As such, the positive monotonicity assumption holds. This is not to say that oocyte quality should be disregarded, rather that both quality and quantity are relevant for the cycle success, given that only collected and fertilized oocytes have the chance, by definition, to develop into a blastocyst stage embryo (Maggiulli et al., 2020; Vaiarelli et al., 2020). Therefore an equilibrium must be sought by defining an optimal number of oocytes to be achieved. For assumption 2 (known optimal outcome), clinicians select the first dose of FSH in order to obtain an optimal number of mature oocytes that is known for all patients, although there is some discussion around what constitutes an optimal number in literature (Sesh Kamal Sunkara et al., 2011; N. P. Polyzos and S. K. Sunkara, 2015; Steward et al., 2014; Ji et al., 2013; Chen et al., 2017). In this project, we have defined it to be between 10 ( $y_{min}^*$ ) and 15 ( $y_{max}^*$ ) mature oocytes, following the recommendations by Sesh Kamal Sunkara et al., 2011 and Steward et al., 2014. This holds true for every patient, even though some will have a reduced ovarian reserve, and will thus not be able to arrive at this range. For these patients, the dose will be adjusted as needed to bring them as close as possible to the optimal range. To note: we have described dose as  $d_i \in [0, \infty)$ , hence there is a minimum set by definition. But this minimum can be higher than 0, and it is very



likely that a maximum limitation exists. Therefore, it is not only physiologically difficult for some patients to get to the optimal outcome range, patients can also be limited by the range of available doses depending on the use case. In this specific use case,  $d_{min}$  has been set at 100 IU of FSH, and  $d_{max}$  ranging from 300 to 450 IU has been explored. Additionally, dose recommendations by IDoser will be transformed from its raw continuous form to a discretized space where increments in dose are done by steps of 12.5 IU. This is done to account for the real available dose steps for FSH administration.

#### 4.4.1 IDoser for FSH dosing

There are two elements essential for the application of our proposed IDoser in all cases: the core model and the loss function. For this study, the core dosing model selected assumes an underlying linear dose-response and is defined as

$$y_i = y_0 + \beta^T x_i d. \quad (4.7)$$

Then, given a desired outcome  $y^*$ , it can be rearranged into our dosing model  $\hat{d}_i$  as follows

$$\hat{d}_i = \frac{y^* - y_0}{\beta^T x_i}. \quad (4.8)$$

This can be generalized to the following dosing model

$$\hat{d}_i = \frac{\kappa}{\beta^T x_i}, \quad (4.9)$$

which will have as parameter set  $\gamma$  both  $\kappa$  and  $\beta$ , that is  $\gamma = (\kappa, \beta)$ .

For the loss function, additional rules (outside the basic ones described) were defined to ensure an improved but conservative dosing policy, as highly variable doses are discouraged due to greater uncertainty regarding the expected outcome. Limitations in dose changes were defined depending on the outcome range for the specific patient. Following the definitions by N. P. Polyzos and S. K. Sunkara, 2015, the next categories were defined:

- An outcome below than 4 mature oocytes was considered too low;
- An outcome between 4 and 9 mature oocytes was considered sub-optimal;
- An outcome between 10 and 15 mature oocytes was considered optimal; and
- An outcome greater than 15 mature oocytes was considered too high.

Accordingly, higher changes were allowed for patients with a too low and too high outcome compared to those with a sub-optimal outcome. Specifically, a dose modification up to 150 IU was

allowed in the first two instances, and up to 75 IU for those in the latter cases. Changes up to 25 IU (two times 12.5, the available step after discretization) were not considered as such. All these thresholds were established in collaboration with expert professionals in the field.

## 4.5 Evaluation Methodology

### 4.5.1 Literature benchmark

We identified from the existing literature the implementation described in the study by La Marca et al., 2012 and later tested via an RCT (Allegra et al., 2017). This work uses a core model similar to the one in our research, ensuring positive monotonicity. Additionally, our second assumption was referenced in the paper by fixing  $y^*$  to 9 oocytes for all patients, leaving implicitly our concept  $y_0$  equal to 0. Thus, we decided to use it as the literature *benchmark* for our study, and from here onward will be referred to as La Marca or LM.

Their covariates included age, AMH and FSH. The developed and published model was derived from running a linear regression of the following equation:

$$\frac{y_i}{d_i} = \beta^T x_i, \quad (4.10)$$

where the coefficients included in  $\beta$  were estimated to construct the dosing model. The dosing model constructed then would be expressed as

$$\hat{d}_i = \frac{y^*}{\beta^T x_i}. \quad (4.11)$$

In the following RCT (ibid.) a significantly higher proportion of patients got an optimal outcome (described here as 8 to 14 oocytes), even if the mean number of oocytes was not significantly changed.

### 4.5.2 Optimization exploration

Several approaches were explored when optimizing the LM model, and compared statistically to clinical practice and the unmodified LM model. Specific details and results are covered in Appendix B. The final model was obtained after including two extra covariates available in our dataset (AFC and BMI), and omitting the variable basal FSH. All parameters of  $\gamma$  were optimized and used for the final proposed doser. A second optimization was run after  $\gamma^*$  was found in order to find a value of  $d_{max}$  across the available dosing space that would minimize  $L$  even if the maximum value for  $d_{max}$  (450) was always allowed. This second optimization where, as in the first one, the

loss function penalizes drastic dose changes, was done to obtain a conservative dosing model even if no maximum dose boundaries were used.

### 4.5.3 Model comparison and statistic tests

To compare the optimized models to LM and clinical practice two methods were used. The first one was analyzing and plotting  $L$  of all options across all the  $d_{max}$  values allowed. In this first method, an extra comparison was performed to understand the quality of the IDoser and LM models. This extra comparison introduced an oracle decision policy or model, where the doses recommended are always correct: if a dose change is needed it is done in the right direction and inside the adequate range. This is done by simply determining if the outcome  $y_i$  is inside or outside the optimal range. If it is outside, a dose change is needed, and it is executed in the correct direction and range. Hence, the oracle model's dose recommendations represent all available and correct dose changes for the test patients, or in other words, a perfect policy as per our loss function. The  $L$  value for the oracle indicates the maximum improvement possible in the validation dataset.

To test if any of the methods were statistically different from clinical practice or among themselves, loss ( $l$ ) values from every group were compared between them. This is a comparison of more than three sets (or groups) of related values (as each patient  $p_i$  has a  $l_i$  value for each method tested). If the data was normally or close to normally distributed, a repeated measures ANOVA test could be used. Normality was tested in our data via a Shapiro test, and was rejected (p-value  $<0.01$ ). Thus, a non parametric statistical test was needed. We used the method recommended in the studies by García and Herrera, 2008 and by García, Fernández, et al., 2010, which are an extension of the study by Demšar, 2006. Specifically, Iman-Davenport's corrected Friedman test (Iman and Davenport, 1980) was used. When significant results were achieved (meaning that significant differences were found between the groups), a post-hoc test was used to determine which groups were different. P-values were adjusted using Finner's correction, as per García, Fernández, et al., 2010. R's package *scmamp* was used to run the mentioned tests. A p-value of less than 0.05 was considered significant.

## 4.6 Results

As mentioned, the first methodology used to compare our proposed model, IDoser, to LM and clinical practice was to plot  $L$  of every model after they were used to dose our validation dataset across all 4  $d_{max}$  explored. Clinical practice was always described always as  $L = 0$ . The values obtained were plotted in Figure 4.4, together with the oracle's value ( $\hat{d}$  always in right direction and range).

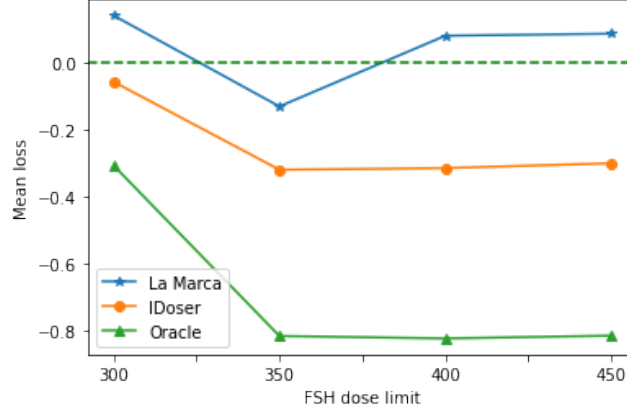


Figure 4.4:  $L$  across  $d_{max}$  for La Marca, oracle and the proposed IDoser when used to dose in the validation dataset. The dashed line marks  $L$  for the clinical practice dosing method.

As it can be clearly observed, IDoser is always under LM  $L$  values and below the 0 mark, where clinical practice lies. It is also clear that the oracle model lies far below both of them, indicating that there is still a gap to be filled.

Regarding statistical results, Iman-Davenport’s corrected Friedman test results (shown in Table 4.2) prove a significant difference between models’  $L$  across all selected points of  $d_{max}$ .

$d_{max}$	300	350	400	450
p-value	0.002657*	4.977e-11*	5.373e-14*	8.36e-14*

Table 4.2: Results of Iman Davenport’s correction of Friedman’s rank sum test of all methods tested across the 4 selected values for  $d_{max}$

The consequent post-hoc test results to ascertain which specific models were different showed a significant improvement of our optimized model compared to the LM model across all  $d_{max}$  points explored. This also holds true when compared to clinical practice, except in the case of  $d_{max} = 300$ , where even if an improvement is observed, it cannot be proven statistically. These results are represented in Table 4.3. Specific  $L$  values and adjusted p-values are listed in Tables B.1 to B.4 in Appendix B.

$d_{max}$	Ordered results by significant differences
300	La Marca $\prec$ Clinical Practice $\sim$ IDoser
350	Clinical Practice $\prec$ La Marca $\prec$ IDoser
400 450	La Marca $\sim$ Clinical Practice $\prec$ IDoser

Table 4.3: Ordered results from worst (left) to best (right) method in one vs one comparison across all  $d_{max}$  values. Results extracted from post-hoc test with p-values adjusted by Finner’s methodology.

## 4.7 Discussion

The proposed IDoser model achieved a significant improvement compared to the LM model across all investigated  $d_{max}$  values, and most of the times also when compared to baseline clinical practice or policy, except for  $d_{max} = 300$ . Here, the improvement achieved did not reach the significance threshold stipulated. As shown on Figure 4.4, there is actually less margin for improvement in dosing policy compared to the rest of values of  $d_{max}$ . This would explain why, with our current sample size, a significant difference compared to clinical practice with  $d_{max} = 300$  cannot be shown, as there are few cases that can be improved, and IDoser does not identify a correct dose change for all of them. This is clearly evident from the distribution of improvable cases across the outcome axis (number of mature oocytes). We define case as improvable whenever  $y_i$  is outside of the optimal range, and  $\hat{d}_i$  can be changed in the right direction. In Figure 4.5 we can observe that most of them are concentrated in low or sub-optimal ranges .

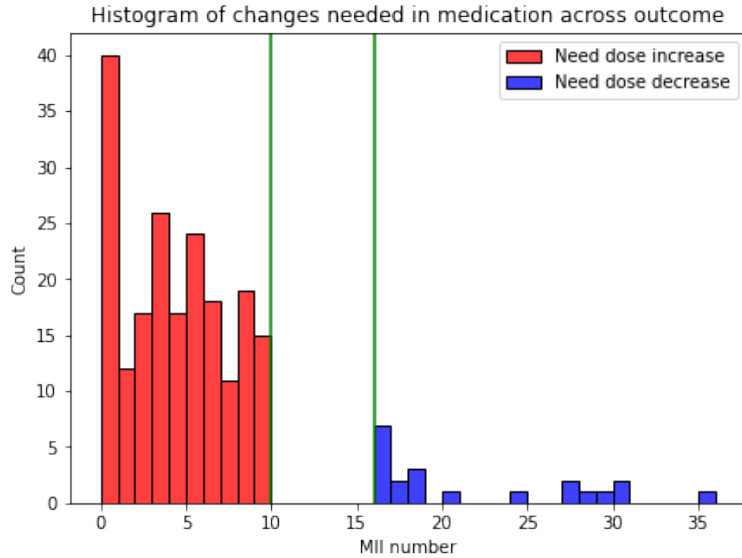


Figure 4.5: Distribution of cases that need an increase of dose (red) or decrease (blue) for the validation dataset if  $d_{max} = 450$  is allowed.

These cases would need a significant increase of medication, however, as expected, many low-responder patients have already received 300 IU of FSH by their clinician. This value of  $d_{max}$  is commonly used in European countries, supported by the new European Society of Human Reproduction and Embryology (ESHRE) guidelines for ovarian stimulation The ESHRE Guideline Group on Ovarian Stimulation et al., 2020. As shown in Figure 4.6, there are still some cases in the validation database that have been dosed over 300 IU in clinical practice, indicating that clinicians thought that some specific patients may benefit from exceeding the broadly recommended  $d_{max}$ , as studies used for the guideline are based on population tendencies and not individuals. Given that the IDoser model has been optimized to also auto bound itself for a more conservative dosing

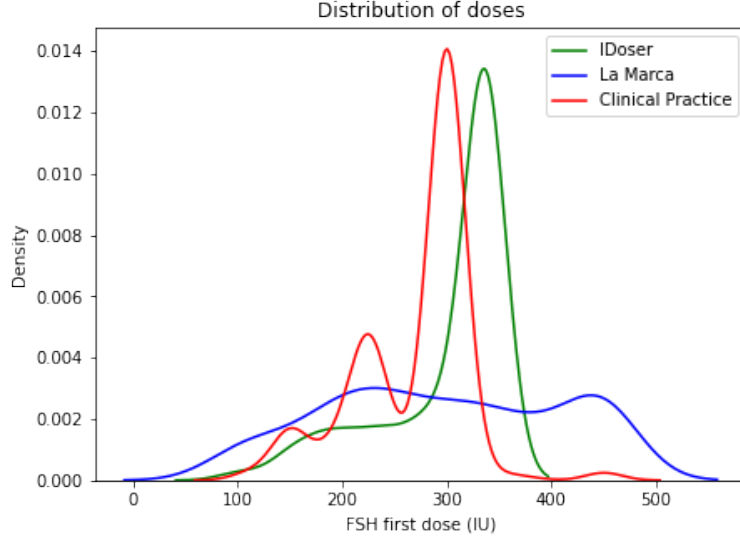


Figure 4.6: Distribution of doses for La Marca (blue), Clinical Practice (red) and IDoser (green) for the validation dataset with  $d_{max} = 450$

policy (with an optimized  $d_{max}$  of 333 IU), it could be used safely with an open  $d_{max}$  in order to identify which patients are candidates for a FSH dose over 300 IU. Figure 4.6 also aids to visualize how IDoser smooths and shifts the dose distribution slightly upwards compared to clinical practice, and how it is limited to 333 IU by its automatic bounding system. On the contrary, the LM model tends to distribute doses more evenly (not as centered in 300), having more cases with decreased doses and doses over 300 concentrated in the 450 IU mark ( $d_{max}$  value implemented in Figure 4.6).

These antagonistic tendencies can also be clearly visualized in Figures 4.7 to 4.10, where dose changes distributed across the outcome are shown for both models (IDoser and LM) and  $d_{max}$  300 and 450. IDoser (Figures 4.7 and 4.9) tends to rescue more cases under our defined  $y_{min}^*$ , where the majority of improvable cases lie, at the expense of very few patients over  $y_{max}^*$  having their dose increased and some not decreased. This could be due to an under-representation of this subset of patients in our dataset, and should be considered a limitation of the resulting model. On the contrary, the LM model performance (Figures 4.8 and 4.10) shows that more patients that need a reduction on dose get it, but at the same time many patients that need an increase in dose are instead given a reduced one. That is clearly why our loss function is penalizing this model.

It could be argued that the decreasing tendency of the LM model could be derived by its use of a  $y^*$  value of 9, one step lower than our defined  $y_{min}^*$  of 10. Of course, this could be the case for some patients in our databases, but as shown in Figures 4.8 and 4.10, it would not explain the relevant dose reductions in patients with very low outcomes. Another cause of the lower performance of the LM model could be that it was originally developed for normo-ovulatory patients under 40 years, and our database comprises all patients eligible for an IVF treatment. That characteristic

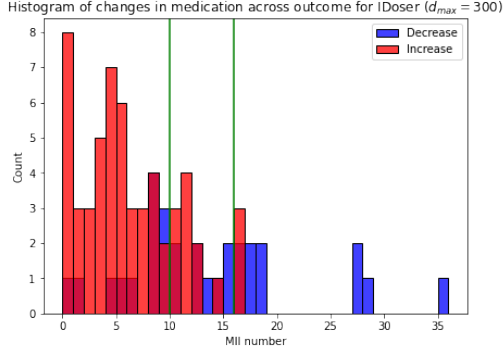


Figure 4.7: Doses changes for the IDoser model with  $d_{max} = 300$

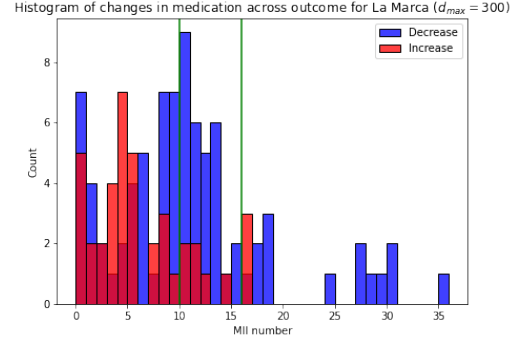


Figure 4.8: Doses changes for the La Marca model with  $d_{max} = 300$

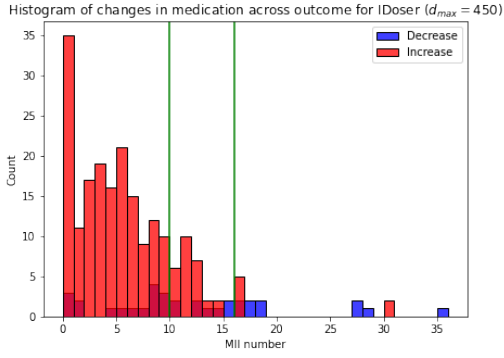


Figure 4.9: Doses changes for the IDoser model with  $d_{max} = 450$

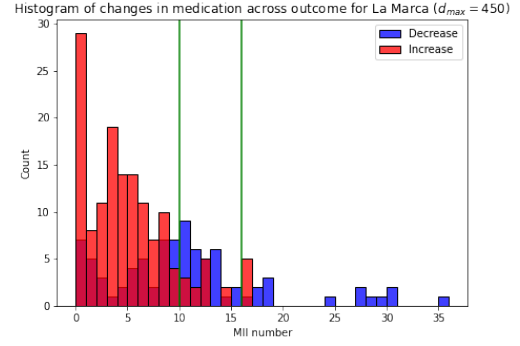


Figure 4.10: Doses changes for the La Marca model with  $d_{max} = 450$

was the motivation for this study, as all published FSH dosing models excluded critical portions of the patient population. Also notable is the vantage obtained with IDoser while not using a single point for  $y^*$ , but a range of desired outcomes. This makes the model less prone to change doses in the desired interval, as shown in Figures 4.7 to 4.10.

Finally, it is also worth noting that our loss function penalizes dose changes considered to be too large, even though they may be in the right direction. Given that our model has been optimized with these rules to avoid being too bold, the fact that it still recommends significant changes in dose could be due to some patients truly needing such a change. Importantly, true clinical utility of IDoser can only be established through a prospective randomized trial.

The methodology described here may be applied to any similar dosing problem. Existing methodologies, such as causal inference and double machine learning are very interesting and robust, but in cases like the one described in this study, far from applicable. Nevertheless, there is still a need to improve dosing policies using evidence-based knowledge for the sake of patients that receive subpar clinical care following generalized policies.

Observational datasets available in clinical practice, generated using dosing protocols, are prone to

have little variability. Moreover, not all confounders can be accounted for evenly across all cases, as different clinicians favor different biomarkers or prescribe more or less complete test batteries depending on experience or other factors, including financial constraints or patient request. Hence, algorithms trained on those databases may not conform with evidence-based knowledge, and even sometimes plain logic. It is clear that formalizing the rules that a clinician applies based on their experience, and packaging them in a selected core dosing model and a loss function can help in the process of obtaining a model that follows those rules. These two rather simple elements allow for significant versatility for implementation in different dosing settings (local and/or negative monotonicity, changing core function, etc.). These concepts partially take inspiration from PK/PD modelling, where physiological and pharmacological assumptions and principles are followed. This is translated into our methodology by including a core model where the monotonic assumption is heeded, and by including rules in our loss function that penalize any dose change in the wrong direction. In this study, we explore a core function derived from a linear dose-response. Other functions, including exponential and a straight linear function relating dose to covariates were investigated, however the one presented gave the best results. Other core models that are closer to physiological dose-response relationships, such as sigmoid functions may be explored in the future. To note, IDoser in its present form is only applicable to single-dose dosing cases, and does not take into account adjustment of dose over time for each individual patients.

Ultimately, IDoser achieved an optimized dosing policy in a time-efficient manner, but can also be implemented in a conservative way to validate drug doses “in silico”. This is especially important, given that RCTs entail a significant investment both in time and money. Being able to demonstrate some expected improvement non-interventionally should lead to a faster approval by appointed regulatory authorities in the route to an RCT.

IDoser constitutes a clear and straightforward method to implement field knowledge to train individualized dosing models with a relevant predicted improvement on current clinical practices. This is especially relevant, as in several instances the historical databases available are not amenable to more complex methodologies.

Future lines of work include the implementation of individual values for  $y^*$ , the use of different optimization methodologies outside of coordinated descent, or more complex core models close to real dose-response functions.



## 4.8 Conclusions

At the start of this chapter, we established as the main objective to address the second research question of this project:

**Q2:** *“Can we extend this methodology to other dosing problems?”*.

IDoser has been described and tested successfully in the FSH dosing case. IDoser can be described as an straightforward method to improve individualized drug dosing policies using available observational datasets and field knowledge, while simultaneously incorporating requirements of the specific problem. Its generalized set-up allows for its extension to similar dosing problems, where monotonicity (positive or negative) and known outcome objective can be assumed. Furthermore, the loss function is highly customizable, which enables the inclusion of multiple requirements regarded as important in any given use case, and can also be modified to consider different reward/penalization weights depending on the dose change made. Along this line, the loss function is the general version of the ad hoc score function presented in Chapter 3, and its inclusion in the optimization phase is reminiscent of the first iteration presented in that chapter. Hence, ideas from both iterations (including the dose evaluation in the training phase, and specification of a core dosing model) have been included in the IDoser method, and specified to allow for its application to similar dosing problems. Future work should focus on broadening its applicability and performance.

Specifically to test its applicability in another setting inside the processes of IVF, Chapter 7 contains an initial approach to the extension of IDoser to optimize the selection of the number of embryos for transfer.

Regardless of the good performance demonstrated “in silico” for the FSH case, prospective validation is still required for iDoser clinical use. In Chapter 5 we will outline a detailed protocol for an RCT to test its non-inferiority against standard clinical practice.

## Chapter 5

# IDoserFSH: A non-inferiority study protocol for a multi-center randomized c trial

In this chapter, we will present a detailed protocol for a randomized controlled trial (RCT) designed to test the non-inferiority of IDoser for FSH (or IDoserFSH) compared to standard clinical practice. This is a slightly modified version of an article currently under revision by the journal Trials.

### 5.1 Background

AI-powered solutions, particularly in the healthcare industry, must be utilized cautiously. It is essential to conduct a thorough evaluation of the models' performance to prevent any harm to patients and ensure that the intervention is at least equal in terms of outcome to the current standard of care, with an ideal improvement on the leading outcome. In clinical settings, RCTs are considered the benchmark for assessing the efficacy of a treatment. Although there are differing viewpoints regarding the need for RCTs in all situations, it is an undeniable fact that current regulatory norms (specially in the EU) are stringent and require prospective randomized validation for any AI model that has an impact on treatment prescription in order for appointed authorities to consider their efficacy and safety sufficiently proven. Consequently, the requirement for prospective validation poses a significant challenge that must be addressed before clinical implementation can take place. There is no exception for FSH dosing models, such as IDoserFSH, the one developed in Chapter 4 of this thesis.

Some of the FSH dosing models available in the current literature have been validated in RCTs

(Allegra et al., 2017; Olivennes et al., 2015; Nyboe Andersen et al., 2017), demonstrating clinical improvements such as an OHSS risk reduction, or an increase of patients achieving the targeted response (8-14 oocytes). Nevertheless, no differences in pregnancy or live birth were confirmed. Furthermore, none of these models consider patients over the age of 40 years or non-normo-ovulatory women in their development nor in the respective RCTs. As such, the applicability of current models remains limited. The trial presented in this chapter evaluates an all-inclusive, comprehensive FSH dosing model for COH (IDoserFSH). This kind of trial for interventional ML models is sparse in literature, but necessary for a safe clinical application of any medical device of this nature.

This chapter is composed by the answers to the SPIRIT checklist (Standard Protocol Items: Recommendations for Interventional Trials; Chan, Tetzlaff, Gøtzsche, et al., 2013; Chan, Tetzlaff, Altman, et al., 2013). This checklist and its accompanying statement, were elaborated by an international group of stakeholders (SPIRIT group) with the objective of improving the completeness and quality of clinical trial protocols. Their recommendations are evidence-based, and developed using a systematic methodology. The checklist adheres to the ethical principles stated in the Declaration of Helsinki (2008), requirements from trial registration by the World Health Organization (WHO) and the International Committee of Medical Journal Editors.

## **5.2 Trial general details**

### **❖ Objectives**

This RCT tests the hypothesis that the performance of IDoserFSH in prescribing the initial dose of FSH during COH is not inferior to the performance of a clinician, measured by the average number of mature oocytes retrieved after COH. Secondary objectives include the effect of IDoserFSH on number of cycle cancellations (due to low response or no retrieval of oocytes), number of cases with OHSS risk, clinical pregnancy and live birth after the first embryo transfer.

### **❖ Trial design**

Single blinded RCT with two arms with a 1:1 allocation ratio, comparing outcomes following FSH dose selection by IDoserFSH and dosing based on standard clinical practice.

The hypothesis that will be tested is whether the performance of IDoserFSH selecting the first dose of FSH is non-inferior to the standard clinical protocol in regards to the average number of mature oocytes recovered per group. The difference in number of MII that is regarded to be clinically significant is 2 MII. If the average number of MII in the intervention group is inferior in 2 or more units, IDoserFSH will be considered inferior to the standard clinical practice.

## ❖ Trial status

Recruitment has not started. Current protocol version is: version 2 from the 2nd of February 2023. The protocol has been reviewed and approved by the Eugin ethical committee.

## ❖ Trial flowchart

The general flow of the study can be visualized in Figure 5.1, with each step in the chart described in detail the following sections.

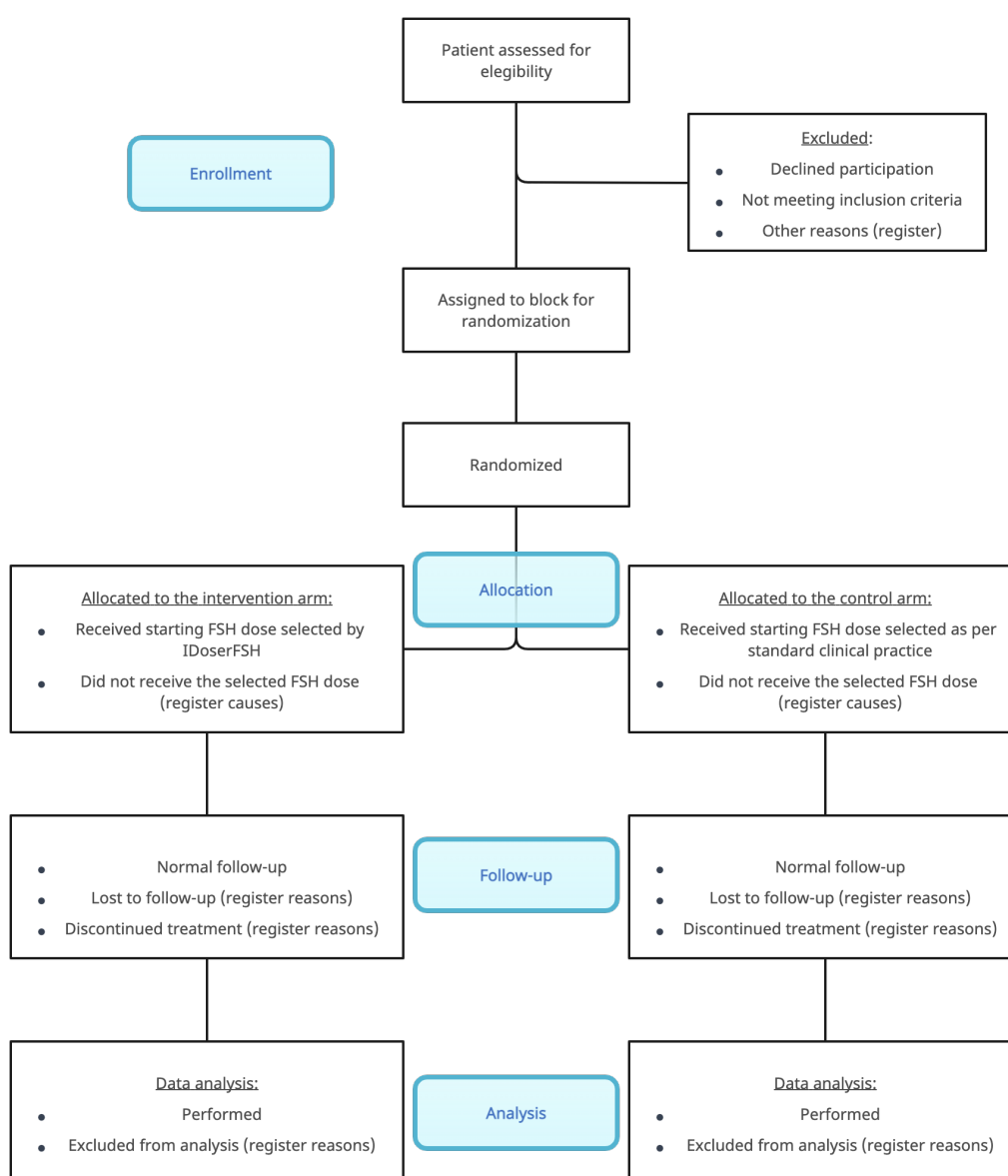


Figure 5.1: Trial step-by-step flowchart.

## 5.3 Methods: Participants, interventions and outcomes

### ❖ Study setting

This is a randomized, single-blinded, multicenter clinical trial.

### ❖ Eligibility criteria

Patients will be selected based on eligibility criteria by their clinician during the pre-treatment visit. Inclusion criteria are: first IVF cycles; use of autologous oocytes (those pertaining to the same patient); and use of FSH on the first day of stimulation (which can be combined with luteinizing hormone, LH). Exclusion criteria are: natural cycles (without COH); and cycles in which FSH is not measured in International Units (IU).

### ❖ Who will take informed consent?

Patients will only be included in the trial after written informed consent is retrieved by medical personnel prior to starting COH. This medical personnel will be composed by the main clinicians assigned to each patient, that follow them up during their treatment at the clinics.

### ❖ Additional consent provisions for collection and use of participant data and biological specimens

Not applicable

## 5.4 Interventions

### ❖ Explanation for the choice of comparators

Patients in the control group will be prescribed their first dose of FSH by the clinician in accordance with standard clinical practice.

### ❖ Intervention description

Patients in the intervention arm will be prescribed the first dose of FSH by IDoserFSH, that will take into account the age of the patient, BMI, AFC and AMH. These data will be retrieved from the patient clinical file after their first visit to the clinic, after providing the patient with informed consent documentation.

#### ❖ **Criteria for discontinuing or modifying allocated interventions**

Participants can withdraw from the study at any time and for any reason, or no reason. The reason for withdrawal will be recorded if patients choose to disclose this information. Failure to administer the allocated FSH dose and discontinuation of IVF for medical reasons will also result in withdrawal. Participants will be communicated about their withdrawal (if not decided by them) as soon as either error of dose administration or a medical reason are detected. If a medical reason or an error in FSH administration is detected, the participant's treatment will either be interrupted or continue as routine practice, depending on the clinician judgement. The data of withdrawn patients obtained during their participation in the study will be included in the study analysis. Withdrawn patients will not be replaced.

#### ❖ **Strategies to improve adherence to interventions**

The research team will ensure that the patient has adequate follow-up after the end of the intervention in order to retrieve all relevant outcome data, which is the standard after treatment for all IVF patients.

#### ❖ **Relevant concomitant care permitted or prohibited during the trial**

Outside of first FSH dose allocation, the COH and IVF treatments will be under control of the assigned clinician of the participant, as per routine practice.

#### ❖ **Provisions for post-trial care**

Care post-trial will follow routine practice and will be controlled by the assigned clinician for each participant.

In accordance to the Spanish legislation regarding clinical trials with medicines (Real Decreto 223/2004 6th of July), the sponsor of the study will subscribe an insurance covering the sponsor, investigator, collaborators and center. This will cover any contingencies in the event of deleterious consequences for participants.

#### ❖ **Outcomes**

The primary efficacy criterion will be the number of mature, MII oocytes retrieved at OPU. This treatment outcome is the closest to the intervention and has a clear impact on IVF cycle success. Secondary efficacy endpoints will include cycle cancellations (due to poor response or in the event that no mature oocytes are retrieved), OHSS risk, clinical pregnancy and live birth per first transfer.

The following variables will be analyzed:

- COH outcome variables
  - Number of MII oocytes
  - Number of cumulus-oocyte complexes (COCs)
  - Estradiol at last ultrasound assessment (pg/ml)
  - Number of follicles  $\geq 11$ mm at last ultrasound check
  - OHSS risk rate (risk = estradiol  $> 5000$  pg/ml and or  $\geq 18$  follicles  $\geq 11$ mm at last ultrasound check)
  - OPU cancellation rate (number of patients who have stopped COH prior to OPU/total number of patients)
  - Cycle cancellation rate (number of patients with no MII oocytes at OPU/total number of patients)
- Pregnancy outcome variables
  - Clinical pregnancy rate (fetal heart beat observed at 7th week of gestation) per first embryo transfer
  - Live Birth rate per first embryo transfer

#### ❖ Participant timeline

Participant schedule for enrolment, interventions, assessments and trial relevant visits can be visualized in Figure 5.2.

A detailed description of the timepoints reflected in the figure is as follows:

- $-t_1$ : First appointment with the clinician where information on IVF treatment is relayed to the patient. Information on the trial is communicated to the patient, and informed consent documentation is handed over to eligible patients. If any of the baseline variables needed for IDoserFSH to function are not available at this time, steps are put in place to obtain relevant information for the next appointment. These can include petition of blood tests for AMH level results, or echography for AFC assessment.
- $t_0$ : Appointment with the clinician where the signed informed consent form is retrieved. Patient is included if all baseline variables are available. Once included, the participant is allocated to either the IDoser FSH selection group or control group. The FSH dose is prescribed to the participant.
- $t_1$ : First day of the COH protocol. Participant will self-administer the prescribed dose of FSH.

TIMEPOINT	STUDY PERIOD						
	Enrolment	Allocation	Post-allocation				Close-out
	- $t_1$	0	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$
<b>ENROLMENT:</b>							
Eligibility screen	X						
Informed consent	X						
Allocation		X					
<b>INTERVENTIONS:</b>							
FSH dose selection by model			X				
FSH dose selection by clinician			X				
<b>ASSESSMENTS:</b>							
Baseline variables	X	X					
Outcome variables				X	X	X	X

Figure 5.2: SPIRIT flowchart of enrolments and assessments: Detailed timing for relevant events for participants during the randomized trial

- $t_2$ : Last ultrasound appointment prior to OPU. First outcome variables are registered (OHSS risk and OPU cancellation).
- $t_3$ : Day of OPU. Further outcome variables are registered (number of COCs recovered, number of MII recovered and cancellation after OPU).
- $t_4$ : Appointment to evaluate the presence of a fetal heart beat at the 7th week of gestation to establish whether pregnancy is achieved after first embryo transfer in current IVF cycle. Presence or absence of clinical pregnancy is registered.
- $t_5$ : End of study, considered after outcome on live birth after first embryo transfer is obtained. Presence or absence of live birth achieved is registered.

Baseline variables include all necessary variables required for IDoserFSH, including age of the patient, BMI, AFC, and AMH levels.

#### ❖ Sample size

A sample size calculation was performed to establish the number of participants required for the study. To determine a statistically significant difference equal or greater to 2 MII oocytes, 118 subjects are necessary in each group ( $n=236$ ), accepting an alpha risk of 0.05 and a beta risk of 0.2 in a one-sided test. We estimate a mean common standard deviation of 5.84 (as per the observational data included during development and validation of IDoserFSH) and anticipate a drop-out rate of 10%.



### ❖ Recruitment

All potential participants will be informed about the study by their clinician prior to undergoing IVF treatment.

## 5.5 Assignment of interventions: allocation

### ❖ Sequence generation

Stratified block randomization will be carried out depending on whether the patient is an expected poor responder ( $AMH < 1.2$  ng/ml and  $AFC < 5$ , as per POSEIDON criteria Esteves, Roque, et al., 2018), high responder ( $AMH \geq 3$  ng/ml and  $AFC \geq 15$ ) or expected normo responder (all other cases). This will ensure equal distribution of these patient etiologies across both arms. There will be 3 blocks (one for each strata) for every arm. The size of every block has been determined by the population distribution of each mentioned strata during the development and validation of the IDoserFSH to be studied (13% poor, 15% high and 72% normo-responders) (Correa, Cerquides, Vassena, et al., 2023). These figures translate to 15 patients in the poor responders group, 18 in the high responders, and 85 in the normo-responders within each arm. The study group (treatment or control) will be randomly assigned using a computer program with a 1:1 allocation ratio.

The 3 randomization lists will be generated using the online software Graphpad<sup>1</sup>. This is a single-blind trial, in which the patients are blinded to the source of the dose prescribed, as are also other clinicians apart from the assigned one to the participant, embryologists and part of the research team.

### ❖ Concealment mechanism

Single-blind trial. Participants will not be aware of which arm they have been placed in. Once the random allocation sequences are generated, they will be stored in an electronic data table accessible to the responsible clinician. Once the data for IDoserFSH is introduced into the table, the arm allocation field will be populated with the next free sequence value of the participant strata. Apart from the data coordinator, the research team will also be blinded to participant allocation.

### ❖ Implementation

The allocation sequence will be generated with the Graphpad tool by a member of the research team. Enrollment will be carried out by medical doctors. Once included, participants will receive the prescription for the first FSH injection dose, whether decided by the clinician or IDoserFSH, and self-administer it as indicated by their clinician.

---

<sup>1</sup><http://www.graphpad.com/quickcalcs/randMenu/>

## **5.6 Assignment of interventions: Blinding**

### **❖ Who will be blinded**

Participants will be blinded to arm allocation. The research team will also be blinded, with the exception of the data coordinator. The medical team, with the exception of the ones assigned to care of the participants, will also be blinded to arm allocation. The embryology laboratory team will be blinded as well.

### **❖ Procedure for unblinding if needed**

Unblinding for participants will be permissible after their participation in the trial is ended. Additionally, it will be permissible for patients, the medical and embryology team in the event of a medical event that justifies unblinding.

## **5.7 Data collection and management**

### **❖ Plans for assessment and collection of outcomes**

All trial data will be collected as per standard clinical practices for IVF treatments in the participant clinics.

### **❖ Plans to promote participant retention and complete follow-up**

There are no additional plans for retention of participants.

### **❖ Data management**

Data will be registered both in the electronic data table where allocation is provided and in a separate registry file. In the electronic data base, maximum and minimum value checks will be performed for every variable (if applicable).

### **❖ Confidentiality**

Participants' information will be stored in both an electronic data table and registry file in a secure folder with controlled access only accessible to clinicians and the research team involved in the trial. This folder will be located in a secure server under exclusive control of the sponsor. Both

the data table and registry file will be password protected. All data will be anonymized. Personal data will be protected according to the Regulation (EU) 2016/679 of the European Parliament and the Council of 27th April 2016 on the protection of natural persons in regards to processing of personal data and on the free movement of such data (General Data Protection Regulation or GDPR). In Spain, in addition to the GDPR, the transposition of this regulation to the national “Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales” will apply.

Relevant data will be stored 10 years after the finalization of the trial as per Regulation (EU) 2017/745 from the 5th of April.

**❖ Plans for collection, laboratory evaluation and storage of biological specimens for genetic or molecular analysis in this trial/future use**

Not applicable

## **5.8 Statistical methods**

**❖ Statistical methods for primary and secondary outcomes**

### **Descriptive analysis**

Description of all demographic and result variables included in the trial will be provided overall and per study group (mean, standard deviation or SD, n, %).

### **Univariable analysis**

Differences regarding the number of MII oocytes amongst the groups will be evaluated using a t-student test or Mann-Whitney U test (if the distribution is not normal). These tests will also be used to compare FSH doses and COC number.

Regarding all categorical variables (OHSS risk, OPU cancellation, cycle cancellation, clinical pregnancy, live birth), differences between groups will be evaluated using Pearson’s Chi Squared. Description of adverse events (if any) will be provided by study group.

The efficacy analyses will be performed in accordance with the intention-to-treat principle.

A p-value <0.05 will be considered as statistically significant.

**❖ Interim analyses**

There is no interim analysis planned for this study.

❖ **Methods for additional analyses (e.g. subgroup analyses)**

Subgroup analysis will be carried out for participants predicted to be low, high and normo-responders.

❖ **Methods in analysis to handle protocol non-adherence and any statistical methods to handle missing data**

Non-adherence to the trial protocol would imply that the participant has either not received the intervention planned or has later dropped out of the IVF treatment/follow-up or not adhered to the prescribed protocol outside of the trial intervention. As such, they would be considered as withdrawn from the trial. In either case, data will be analyzed following the intention-to-treat principle.

Missing covariates would immediately be considered as exclusion criteria, as they are necessary for the randomization and/or the use of the IDoserFSH.

❖ **Plans to give access to the full protocol, participant level-data and statistical code**

Fully anonymized participant level data and the statistical code used for this trial will be made available by the corresponding author on reasonable request.

## **5.9 Oversight and monitoring**

❖ **Composition of the coordinating centre and trial steering committee**

This is a multicenter trial, where the coordinating centre is Clinica Eugén in Barcelona. Daily support for the study is provided by the:

- Principal investigator: supervises the trial and coordinates the study team.
- Data coordinator: manages data annotation and data safety and quality
- Research team: includes both principal investigator and data coordinator, together with co-investigators in charge of data outcome analysis.
- Medical team: in charge of participant recruitment, handling of informed consent forms, follow-up of participants, and safety monitoring according to protocol.

There is no steering committee.

#### **❖ Composition of the data monitoring committee, its role and reporting structure**

The unblinded data coordinator will be in charge of monitoring safety and quality of data and will report to the principal investigator. The data coordinator is not independent from the sponsor. The research team has no commercial conflict of interest.

#### **❖ Adverse event reporting and harms**

Clinical study participants will be routinely asked about adverse events (quantity and quality) at each study visit. Any adverse event that may occur to the participants of the study must be documented and followed up by the investigator. The event will be documented with the necessary investigations for adequate assessment of causality as established in the document “MDCG 2020-10/1- Safety reporting in clinical investigations of medical devices under the Regulation (EU) 2017/745”. Serious adverse events must be immediately notified to the sponsor, who will be in charge of reporting the events to the Ethics Committee and the Competent Authorities. The sponsor must report serious adverse events within 15 days (7 days in case of death or a life-threatening event) using the official serious adverse events notification forms. The sponsor will report the serious adverse events through Eudavigilance-CT.

#### **❖ Frequency and plans for auditing trial conduct**

This trial is subject to external audit independent from investigators and the sponsor, annually. The data coordinator will perform internal audits quarterly, by randomly selecting a subset of participants and crosschecking the electronic data table and informed consents.

#### **❖ Plans for communicating important protocol amendments to relevant parties (e.g. trial participants, ethical committees)**

Any amendment to the trial protocol will be reported to the Ethics Committee and Competent Authorities, and only applied after their approval.

#### **❖ Dissemination plans**

Plans to disseminate results and conclusions of the trial include scientific papers and/or congress communications.

## 5.10 Conclusions

In this chapter, a highly detailed protocol for an RCT to validate the non-inferiority of IDoserFSH compared to standard clinical pregnancy was presented. After the conclusion of the study, a more certain determination regarding its appropriateness for clinical implementation can be conducted. If the non-inferiority of IDoserFSH compared to standard clinical practices is demonstrated, regulatory bodies can approve the next steps (like CE marking for medical devices in the European Union) for its clinical implementation.

RCT trials for ML models that intervene in treatment prescription, as already mentioned, are currently sparse in literature. There is a clear struggle between their necessity, especially for high-stakes medical decisions, and their cost both in time and resources. Administrative changes could be achieved upon collaboration between AI experts and regulatory bodies in order to fine tune which type of AI-driven medical solutions call for an RCT to gather enough evidence of their efficacy and safety in the eyes of appointed authorities. However, up until that moment arrives, the utmost care must be applied whenever selecting a candidate for an RCT. This can be achieved by examining exhaustively the in-silico performance of any candidate, and determining if its recommendations are coherent with field-knowledge. IDoser has been developed in order to accomplished precisely this, and when applied for the FSH dose recommendation task, it has been examined in detail before deeming a good candidate for an RCT trial (see Chapter 4).

In the following chapter, a first approximation to the selection of number of embryos for transfer is described, with an initial approach based on the IDoser method later described in Chapter 7.

## **Chapter 6**

# **Limits of conventional Machine Learning methods to predict pregnancy and multiple pregnancy after embryo transfer**

In this chapter we will present a first approximation to the optimization of the selection of the number of embryos for transfer using out-of-the-box ML methods. This chapter is a slightly modified version of the paper presented as an oral communication in the 23rd International Conference of the Catalan Association for Artificial Intelligence (CCIA) as Núria Correa, Rita Vassena, et al. (2021). “Limits of Conventional Machine Learning Methods to Predict Pregnancy and Multiple Pregnancy After Embryo Transfer”. In: CCIA. DOI: 10.3233/faia210141.

### **6.1 Background**

A core objective of this thesis has been the construction of ML models that abide by constraints determined by field knowledge. Without that, trusting recommending models trained from observational data is certainly challenging. This has been the case in optimizing the dosing policy for FSH, and it is not an isolated case. In several healthcare fields, a robust research background already exists, providing a high amount of field knowledge. In this context, it is expected that known data relations are picked up by trained models and their predictions heed them. In other words, the models’ decisions need to be coherent with previously demonstrated knowledge. Furthermore, to ensure user confidence and general transparency, the explainability of ML models is of paramount importance. Being able to explain in a relatively simple way how the models arrive

to their decisions also allows for a better inspection of their adherence to field knowledge. To sum up, expectations on how the models will work are set by preceding research, and failure to comply with it hinders the models' applicability and logically, diminishes the confidence of the users in the models' predictions.

The specific case covered in this chapter, the selection of the number of embryos for transfer in an IVF treatment, is no exception. A brief reminder of the main challenge of this decision follows. After oocyte retrieval and fertilization in vitro with a processed sperm sample, the resulting embryos are cultured in the IVF lab for a few days. Then, selection of those expected to have better chances of giving rise to a healthy pregnancy leads to their transfer to the uterus of the patient. IVF provides approximately 30% pregnancy rate per treatment, which leads to about 20% delivery rate (De Geyter et al., 2018). These rates can undoubtedly be frustrating for both professionals and patients. Until recent years, double embryo transfer (DET) has been standard practice in order to compensate these low success rates. This strategy does increase the rates of pregnancy compared to single embryo transfer (SET), but it also increases the occurrence of multiple pregnancies (Kamath et al., 2020). Compared to singleton births, twin births have higher obstetrical risks (Crosignani et al., 2000). Clinically, repeated SET is the logical solution, as its success rate is equivalent to that of a one-time DET treatment (Kamath et al., 2020), while drastically reducing the chances of a multiple pregnancy. However, lower embryo quality can see the chances of multiple pregnancy reduced, and singleton pregnancy chances raised enough as to reach a risk-benefit scenario that actually calls for a DET as opposed to repeated SET. Additionally, there is a portion of patients that will drop out after a negative treatment, even if there are still embryos that are considered apt for transfer or, in other words, they still have chances to achieve a pregnancy. It is important to remember that often economical and psychological factors play a relevant part in the patients' decisions regarding following their treatment and/or the selection of the number of embryos to get transferred, even if thoroughly counseled by expert clinicians.

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. In order to do so, in this chapter, out-of-the-box ML methods will be explored.

Then, the technical objective is training 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 for those models to be really regarded as usable in clinical practice they need to comply with previously demonstrated knowledge, leading us to identify three main constraints:



1. **Constraint 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 (positive monotonicity). [Kamath et al., 2020]
2. **Constraint 2:** Under any conditions MP chances cannot be higher than P chances.
3. **Constraint 3:** Chances of P and MP are highly correlated with age, embryo stage, and quality (Hardarson et al., 2003; Glujovsky et al., 2016; Grøndahl et al., 2017).

To test the performance of conventional ML models, we need to examine their compliance with all three constraints, further to standard measures such as AUC.

There are several available models focused on predicting the reproductive results of SET treatment versus DET (S. A. Roberts, L. McGowan, et al., 2010; Stephen A. Roberts, Linda McGowan, et al., 2011; Vaegter et al., 2019; Wen et al., 2022). They report their prediction power mainly in AUC score values, but hardly explore explicitly their compliance with field knowledge (the 3 constraints). As such, similar methods are explored in this chapter in order to ascertain if out-of-the-box ML models can comply with them without any specific adaptation. At the time of work, a real clinical dataset was not yet accessible. A synthetic database was constructed based on statistics reported on a real dataset, as described in the next section.

This chapter will be structured in 4 sections:

- **Material and Methods:** Detailing the construction and descriptive analysis of a synthetic database, and the method to examine how models trained on it perform predicting P and MP.
- **Results:** Specifying the prediction scores of trained models and their adherence to the above determined constraints.
- **Discussion:** A comprehensive evaluation of the outcomes obtained in the preceding sections.
- **Conclusions:** Learnings from the work detailed in the chapter are examined and future needs identified.

## 6.2 Material and Methods

There are multiple public and published sources that report results on P and MP with both SET and DET (S. A. Roberts, L. McGowan, et al., 2010; Aldemir et al., 2020). All these population level studies are coherent among them but offer only summarized sample statistics, and no granular

patient level datasets are publicly available. In this chapter, to ensure reproducibility, we focused on the data from the observational study by Aldemir et al., 2020, 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, results on P and results on MP.

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  $33.28 \pm 4.1$  for the first group,  $34.4 \pm 3.8$  for the second and  $29.2 \pm 4.1$  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 Tables 6.1 and 6.2, had less than a 5% deviation compared to the original study results. Further, strict restrictions were put in place in order to avoid inconsistencies in our artificial dataset, such as cases with positive MP results but a negative P result.

	<b>DET with GQEs</b>	<b>DET with MQEs</b>	<b>SET with GQE</b>
	<b>n=324</b>	<b>n=127</b>	<b>n=887</b>
<b>P</b>	41.05	35.43	29.99
<b>MP</b>	23.46	9.45	3.16

Table 6.1: Proportions of pregnancy and multiple pregnancy instances by group in patients who got transferred embryos at the cleavage stage

	<b>DET with GQEs</b>	<b>DET with MQEs</b>	<b>SET with GQE</b>
	<b>n=174</b>	<b>n=52</b>	<b>n=734</b>
<b>P</b>	56.32	25.00	43.46
<b>MP</b>	32.76	23.08	2.32

Table 6.2: Proportions of pregnancy and multiple pregnancy instances by group in patients who got transferred embryos at the blastocyst stage

Three common ML classifiers were selected to be trained on our resulting database: Logistic Regression (LR), Random Forest Classifier (RFC), and Gradient Boosting Classifier (GBC). A randomly selected 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. To do this, all cases

in the test dataset (1) got predicted probabilities of P and MP with SET and DET separately to detect any negative “effects” of increasing the number of embryos; (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.

## 6.3 Results

After analyzing common prediction scores as AUC and accuracy, the LR and GBC classifiers fare better at predicting both outcomes, with LR being slightly better at AUC and GBC at accuracy (see Table 6.3). 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 (Kamath et al., 2020) This is not the case of MP, where multiple RCTs pooled suggest an increase between 11% and 13% if DET strategy is used compared to SET. RFC and GBC are slightly over those values, and LR is very clearly out of the described range.

	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

Table 6.3: Results of the divided by type of model (Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier) and outcome (Pregnancy and Multiple Pregnancy). Mean effect shows the mean differences between chances predicted with DET minus chances predicted with SET.

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 6.1.

Looking upon the second constraint we found no compliance across all models studied, with GBC infringing the constraints the least (see Figure 6.2).

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 6.3. Predicting MP though, GBC seems to show some variation across ages, but LR does not comply at all.

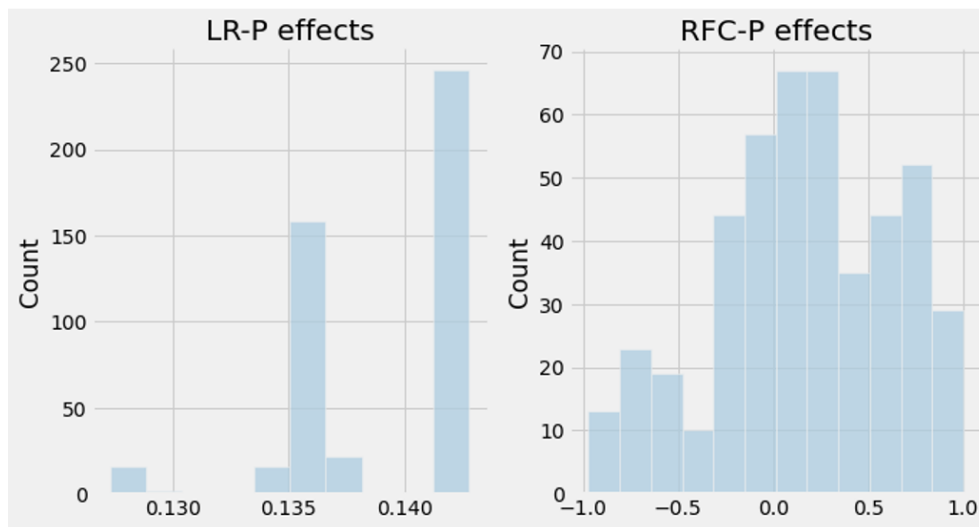


Figure 6.1: Probability differences between predictions on the same patients with SET and DET in the models Logistic Regression (left) and Random Forest Classifier (right) trained to predict pregnancy outcomes.

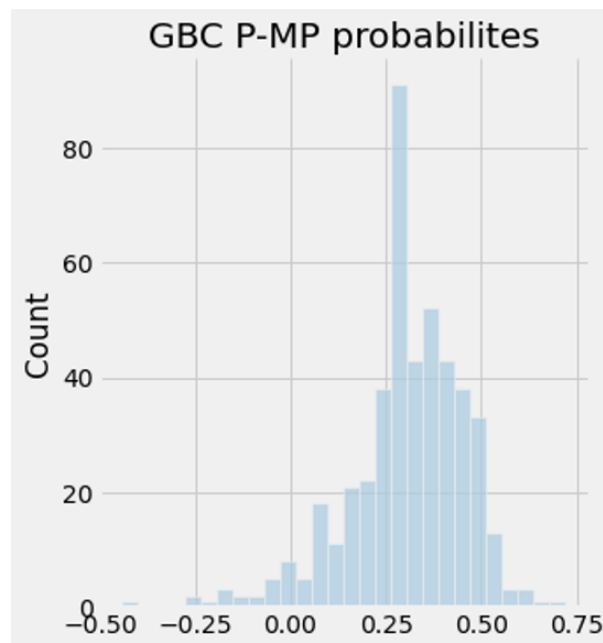


Figure 6.2: Distribution of the differences in predicted probabilities for pregnancy and multiple pregnancy using Gradient Boosting Classifiers.

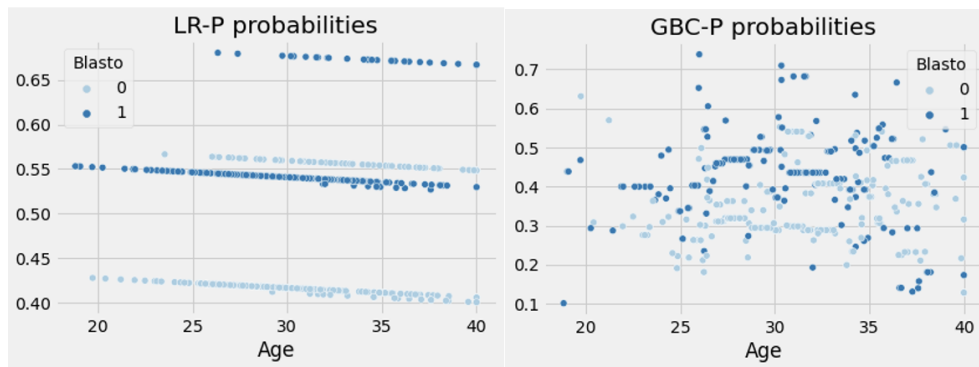


Figure 6.3: Logistic Regression and GBC pregnancy predicted probabilities plotted against maternal age and colored by embryo stage (blastocyst yes or no).

## 6.4 Discussion

Given the poor adherence to the 3 constraints, the three conventional ML models tested are not entirely suitable for the task at hand, even though the AUC scores are close to those available in the medical literature. The highest performing algorithm seems to be LR but it only complies with some of the constraints. If presented to a field expert for clinical practice, it would not be regarded as clinically robust and thus unusable. In any ML implementation related to healthcare, a model needs to be accurate but it also needs to convince the professional about its reliability, i.e. being consistent with field 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 (positive monotonicity), 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 lower quality, nor DET with both embryos of lower 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 remains a concern. Unfortunately, none of the published researches check for that constraint.

As for the second constraint (MP chances cannot exceed P chances in individual cases), one of the main problems 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 classifier 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 (P and MP are correlated to age and to embryo stage and quality) seems to be mostly complied with in published studies. In these publications, the datasets include far more information than the one here constructed, indicating that possessing a database with richer characteristics would enable us to get models compliant with it. To this extent, in Chapter 7, a real clinical database with more relevant variables is used.

## 6.5 Conclusions

In this chapter 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 patient level. In doing so, they fail to comply with clinical knowledge derived constraints. The issue of suitability for population analysis but non applicability for individual level prediction is also reported in S. A. Roberts, L. McGowan, et al., 2010, even if not by explicitly checking adherence to our constraints.

In healthcare specifically, explainability is very relevant as it enables a straightforward analysis of the model's alignment with previous field knowledge. As exposed in other studies (Obermeyer et al., 2019), failing to ensure cohesiveness to it can lead to diminished user confidence in the model and, in the worst-case scenario, detrimental consequences for patients.

Focusing on the specific experiment detailed in this chapter, there seems to be possible solutions for the second and third constraints (**C2**: MP chances cannot be higher than P chances; **C3**: chances of P and MP are highly correlated with age, embryo stage and its quality) , but for the first one (**C1**: chances of P and MP cannot be decreased by increasing the number of embryos to transfer) there seems to be no straightforward answer without adapting the learning process. This exploratory experiment and its results tie with the core concept of this thesis: to obtain clinical robust AI models from biased observational data field knowledge needs to be codified into the training process. The method we called IDoser, described in Chapter 4, is designed to do exactly that. Thus, the natural progression after the experiment presented here, is to apply IDoser to this problem, which would further answer the third research question:

*Can we extend this methodology to other dosing problems?*

Preliminary results of applying IDoser to optimize the selection of the number of embryos for transfer are reviewed in the next chapter, along with an evaluation of the main conclusions and contributions of this thesis and future lines of work.

## Chapter 7

# Conclusions and Future Work

In this chapter we will review the lessons learned from the research in this thesis, the main contributions made and their correspondent publications, and finally, outline the open lines of work to be tackled in the future.

### 7.1 General conclusions

This thesis focuses on the use of AI within clinical environments, particularly in the context of artificial reproduction techniques (ART), with the objective of enhancing the process of one-time dose selection optimization. As it is often the case in such clinical settings, historical data gathered from past treatments constitute the main source of information available. These historical datasets contain information on non-randomized dose allocations, as are derived from day-to-day clinical practice. Consequently, as routine practice is focused on the best interest of the patient, standard procedures are guided by protocols, experience, and the scientific literature on the specific drug and its effect on the desired outcome. This leads to similar patients receiving similar doses, which produces datasets with limited variability in dose allocation. This is also related to the confounding effect found in them, as doses are allocated depending on specific covariates that also affect outcome.

Out-of-the-box ML methods are not suitable to optimize dose selection in these cases, as the information captured in clinical observational datasets is insufficient for them to learn the true underlying dose-response relationship, as explicitly explored in Chapter 6. It is a commonly said that AI learns from what we show it, and this is clearly the case here, where clinical observational datasets lack enough information for the model to learn the proper dose-response relationship.

Pharmacometrics (PX) and causal inference methods are good approaches to model these dose-response relationships. PX is based on mathematical models that already introduce constraints



on how the functions can be modelled, complying with known characteristics of the drug studied. This method, though, works best with prospective randomized data or with varied observational data that includes drug blood levels. Causal inference does work perfectly with observational data, but needs varied enough dose allocation and complete information on all confounders known to affect the dose-response relationship. The variation on dose allocation has been already discussed to be challenging to be obtained in observational clinical datasets. Additionally, obtaining complete information on all the confounders is certainly difficult, as different medical criteria from each practitioner and economical barriers can hinder the test of all relevant variables. Even-though these approaches cannot be implemented in many clinical settings, their logic has been an inspiration for the solutions presented in this thesis.

As models need to be trained on a more complete picture than the limited observational datasets available in clinical settings, it is clear that we need to introduce field knowledge in the process. In this thesis, we explore this in Chapters 3 and 4 by introducing a core dosing model that complies with known characteristics of the drug-outcome function like monotonicity (or even just local monotonicity). This concept is inspired by the PX approach, where specific curves are fitted that comply with certain pharmacological properties. It is further reinforced in Chapter 4 by the introduction of a loss function into the training of the dosing model. This loss function and the ad hoc performance score described in Chapter 3 evaluate dose recommendations obtained from the model by comparing them to real prescribed doses that are associated to a known outcome. This counterfactual evaluation, inspired by the causal inference approach, is done thanks to expert knowledge on the dose-response relationship. To summarize, codification of expert knowledge does not only lead to dosing models that perform better than clinical standard, it enables them to be clinically robust and trustworthy, as they are trained to comply with specific constraints deemed relevant by field experts.

With the ever increasing interest of AI in clinical settings, concerns are being raised about its safety for patients and about the explainability and accountability of models. In order to gain the trust of both practitioners and patients, the models need to be as transparent as possible, and demonstrate their compliance to field knowledge. In the specific case of dose-response optimization, we propose the use of fully interpretable models that align with pharmacological properties of the drug and its causal effect on the relevant clinical outcome.

Additionally, the in-silico evaluation of the models using counterfactuals coded into a loss function or performance score is key for a proper pre-clinical evaluation of dosing models. This is especially relevant in the dosing models context, as interventional RCTs are a needed step before real clinical implementation is allowed. Due to the high cost both in time and economical investment of this type of trial, proper curation of candidate models is needed.

In the next section, the relevant contributions derived from this thesis are listed and briefly explained.

## 7.2 Contributions

In this section, we list and concisely review the contributions to the state of the art described in the previous chapters.

- **Chapter3:**

- **Contribution 1:** An all-patient inclusive FSH dosing model that demonstrates better pre-clinical performance than standard clinical practice.

This contribution answers **Q1** (*"Is it possible to improve clinical FSH dosing policy for Controlled Ovarian Hyperstimulation using only historical data?"*) by training a linear regression model to predict the slope of an assumed linear dose-response function (or ovarian sensitivity) using covariates related to ovarian reserve and relevant demographics (age, BMI and previous fertility). This predicted slope is the input of a dosing formula derived from the linear dose-response function, together with a desired outcome, which then outputs the recommended dose of FSH.

By assuming the FSH to number of mature oocytes dose-response relationship as linear, we are adhering to the hypothesis that the function is monotonic, derived from field knowledge. Adherence to this not only enables the model to perform well, it also results in heightened trust by clinical professionals, as it aligns coherently with their experience. This trust is also inherently tied with the model's high interpretability level, as professionals of the field can understand easily the origin of the dose recommended by reviewing whether the ovarian sensitivity (or slope of the linear dose-response function) predicted for that individual case is high or low, and why is it predicted so by consulting the linear regression model linked to relevant covariates.

Further, even if in this thesis the value for  $y^*$  has been set to a fix value extracted from the clinical literature, it can be changed to any value a clinical expert deems necessary for individual cases. This flexibility allows for a versatile clinical use of the model.

- **Contribution 2:** An ad hoc performance score for FSH doses.

In order to answer **Q1**, its subquestion **Q1a** (*How can we analyze a dosing model's performance before clinical intervention?*) needed to be answered too. The ad hoc performance score presented in this thesis does this by codifying clinical knowledge in order to evaluate counterfactual scenarios where for the same patient, from whom a

known dose-outcome value pair is known, a confident (even if not granularly precise) estimate can be made on whether a different dose will have a beneficial or detrimental effect compared to the real scenario. This allows for a reliable pre-clinical assessment on the performance of the dosing model, facilitating the selection of candidate models for a necessary RCT in the way of clinical implementation.

- **Chapter4:**

- **Contribution 3:** A general methodology for one-time dose optimization with clinical observational datasets (IDoser).

This contribution answers **Q2** (*Can we extend this methodology to other dosing problems?*) by generalizing the key concepts of **Contributions 1 and 2**. The ad hoc performance score is translated into a customizable loss function that is able to evaluate counterfactual dose scenarios based in comparison to ground truth, and also is used to optimize the core dosing function by coordinated descent. This introduces field knowledge in the training step via the loss function and the selection of core dose function. This introduction of expert knowledge during the optimization of the model's coefficient allows to cover the gap left by the lack of necessary information inherently built in many clinical observational datasets, which hampers the application of methodologies already described in literature.

The method we describe, IDoser, needs just two assumptions: first, that there is a known desired outcome to be achieved and second, the monotonicity of the dose-response relationship in study. This monotonicity constraint can be either positive or negative, and does not need to be general, as it can be assumed only locally. Dose recommendations then could be made confidently in the local space where monotonicity can be assumed, but not outside of it. Including the concept of local monotonicity also allows for the addition of a measure of caution, as outside the space assumed as monotonic there is uncertainty about the counterfactual scenarios.

- **Contribution 4:** An FSH dosing model (IDoserFSH), the result of the implementation of **Contribution 3** that performs pre-clinically significantly better than clinical practice and a literature benchmark.

- **Chapter5:**

- **Contribution 5:** A randomized controlled trial (RCT) protocol for final validation of

the proposed IDoserFSH.

As all AI-driven models designed to be used in clinical settings, especially if at an intervention level, need to be properly evaluated both pre-clinically and clinically. As per current regulatory directives in the EU, decision support systems that influence treatment allocation need to be tested via an RCT before its clinical use is approved. IDoserFSH is no exception, and as such, in this thesis a detailed RCT protocol to test its non-inferiority regarding the number of mature oocytes obtained, and as compared to standard clinical practice, is presented.

- **Chapter6:**

- **Contribution 6:** An exploration on the use of out-of-the-box ML methods to predict pregnancy and multiple pregnancy chances.

As a preliminary exploration on the extension of **Contribution 3** to the optimization of the selection of the number of embryos for transfer, a synthetic database was constructed based in literature, and conventional ML methods were applied to predict the chances of both pregnancy and multiple pregnancy depending on embryo quality, stage, and number. Adherence to the main constraints derived from field knowledge was analyzed, and poor compliance was found with the different out-of-the-box ML models tested. This initial investigation proved the importance of introducing field knowledge to the model training process in order to achieve adherence to known constraints.

### 7.3 Publication list

1. Núria Correa, Flavia Rodríguez, et al. (2021). “P-637 Development and validation of an Artificial Intelligence algorithm that matches a clinician ability to select the best follitropin dose for ovarian stimulation”. In: *Human Reproduction* 36.Supplement\_1. deab130.636. DOI: 10.1093/humrep/deab130.636

A poster communication in ESHRE 2021, where an FSH dosing model and its pre-clinical results as compared to clinical practice are presented. The model and ad hoc dose performance score in this publication are described in Chapter 3, section 3.2. At the time of work, only partial validation results were available, leading to find a non-significant difference between the dosing model and standard clinical practice.

2. Núria Correa, Rita Vassena, et al. (2021). “Limits of Conventional Machine Learning Methods to Predict Pregnancy and Multiple Pregnancy After Embryo Transfer”. In: CCIA. DOI: 10.3233/faia210141

An oral communication paper presented at the 23rd International Conference of the Catalan Association for Artificial Intelligence (CCIA) where out-of-the-box ML methods are used to predict the chances of pregnancy and multiple pregnancy depending on the strategy used, single embryo transfer (SET) or double embryo transfer (DET). Using a synthetic database constructed based in literature, field knowledge derived constraints are checked for multiple standard ML algorithms. Results showed that adherence to the defined constraints is very low, concluding that extra measures need to be applied to obtain clinically coherent models. A modified version of this paper can be found in Chapter 6 (**Contribution 6**).

3. Núria Correa, Jesús Cerquides, Josep Lluís Arcos, and Rita Vassena (2022). “Supporting first FSH dosage for ovarian stimulation with machine learning”. In: *Reproductive BioMedicine Online* 45.5, pp. 1039–1045. DOI: 10.1016/j.rbmo.2022.06.010

This paper covers **Contributions 1 and 2**, and describes the development and pre-clinical validation of an all-patient inclusive FSH dosing model. Compared to standard clinical practice using an ad hoc performance score, results showed that the model achieved a significant improvement. Details on this experiment can be seen in Chapter 3, sections 3.2 and 3.3.

4. Núria Correa, Jesús Cerquides, Amelia Rodríguez-Aranda, et al. (2022). “379/427 Acompañamiento en la selección de la dosis de FSH para estimulación ovárica mediante machine learning.” In: *33º Congreso Nacional Sociedad Española de Fertilidad*. Oral communication. Sociedad Española de Fertilidad (SEF). Bilbao

An e-poster communication in 33º Congreso Nacional Sociedad Española de Fertilidad, where **Contributions 1 and 2** were presented.

5. Núria Correa, Jesús Cerquides, Josep Lluís Arcos, and Rita Vassena (2023). “EP-226 Aide à la sélection de la dose de FSH pour la stimulation ovarienne à l’aide du Machine Learning”. In: Oral communication. Pari(s) Santé Femmes. Lille

An oral communication presented in Pari(s) Santé Femme, where **Contributions 1 and 2** were described.

6. Núria Correa, Jesús Cerquides, Rita Vassena, et al. (2023). “IDoser: Improving individualized dosing policies with clinical practice and machine learning”. In: *medRxiv*. DOI: 10.1101/2023.03.28.23287859 (Currently under review by the *Expert Systems With Applications* journal)

This paper introduces a method called IDoser that addresses the limitations of training dosing models using clinical observational datasets. The approach utilizes a customizable loss function, which evaluates counterfactual dose scenarios compared to ground truth, to optimize a core dosing model. Field knowledge is incorporated in the training step through the loss function and the core model. This method is applied to the FSH use case, and compared to a literature benchmark and standard clinical practice. Results showed a significant improvement achieved by applying IDoser. Details on this experiment (**Contributions 3 and 4**) can be found in Chapter 4.

7. Núria Correa, Jesús Cerquides, Josep Lluís Arcos, Rita Vassena, and Mina Popovic (2023a). “O-185 A clinically robust machine learning model for selecting the first FSH dose during controlled ovarian hyperstimulation: incorporating clinical knowledge to the learning process.” In: Oral communication. European Society of Human Reproduction and Embryology (ESHRE) Annual Meeting. Copenhagen

An accepted oral communication to be presented in ESHRE 2023, where **Contributions 3 and 4** are presented to a clinical public.

8. Núria Correa, Jesús Cerquides, Josep Lluís Arcos, Rita Vassena, and Mina Popovic (2023b). “Personalizing the first dose of FSH for IVF patients through machine learning: a non-inferiority study protocol for a multi-center randomized controlled trial”. Under review by the *Trials* journal

A paper describing a detailed RCT protocol to test the non-inferiority of **Contribution 3** (IDoserFSH) regarding the number of mature oocytes obtained, and as compared to standard clinical practice (**Contribution 5**).

## 7.4 Future work

In this thesis, we have described the results of 4 years of research. These results, like the method IDoser presented, and its use-case application IDoserFSH, are not closed ideas. Hence, they can and should be expanded. New avenues to be explored will be described in this section, together with some partial results on the implementation of IDoser for the selection of the number of embryos for transfer.

### 7.4.1 IDoser method

The IDoser method, as presented in this thesis, introduces concepts from both pharmacometrics (PX) and causal inference into the training and evaluation of dosing models, using clinical observational datasets. In the first case (PX) by assuming monotonicity (a mathematical property common in many described dose-response relationships), and in the second (causal inference) by the introduction of the loss function (able to evaluate to some extent the expected effect of counterfactual doses). The core function exposed here is derived from a linear dose-response function, which can be close to reality, but not as close a sigmoid function. Future work in this avenue would include exploration of closer to physiology like the sigmoid one. To do so, is necessary to analyze how the multiple parameters relevant to these function should be linked to pertinent covariates.

Additionally, introduction in the loss function of the saturation effect could be explored. In other words, the loss function could include constraints on increasing the dose whenever the saturation level is estimated to be reached, as more dose is not going to improve the effect, while it is probably going to increment the economical cost of the case and the risks of exposure to the drug.

Exploring these concepts can improve the fit of the dosing models to the physiological and clinical reality, and thus, improve their performance.

### 7.4.2 IDoserFSH

In Chapter 5 a detailed protocol for an RCT to test the non-inferiority regarding the number of mature oocytes retrieved of IDoserFSH as compared to clinical practice is described. This RCT protocol has been already approved by the ethical committee of Eugin (CEIm Eugin) and is currently under review by the clinical journal *Trials*. Additionally, local appointed authorities will need to give their approval for the trial to be able to start.

Whenever this protocol has all necessary authorizations, the trial can start. In the future, analysis of its results will ascertain whether IDoserFSH is indeed non-inferior to standard clinical practice. This prospective comparative clinical evidence would enable the regulatory authorities to approve

its use in clinical practice, which will finally allow for the work reported here to reach its intended use, and hence, benefit both professionals and patients. Reporting these results and their analysis in the form of a scientific paper is also part of the future work.

### **7.4.3 IDoser for selection of number of embryos for transfer**

In order to address the issue of the low success rate per cycle of IVF treatments, it has been common practice for many years to transfer two embryos to the uterus simultaneously. This approach significantly improves the chances of achieving a pregnancy compared to a Single Embryo Transfer (SET) (Kamath et al., 2020). However, this increase in success rates also comes with a higher risk in terms of obstetrical complications, as indicated by the concerning 17% rate of twin births resulting from DET. Twin births carry a four times higher risk of perinatal mortality compared to singleton births. Additionally, twin pregnancies are associated with an increased risk of obstetric complications, such as miscarriage, pregnancy-induced hypertension, gestational diabetes, premature labor, and abnormal delivery, when compared to singleton pregnancies (Crosignani et al., 2000). Consequently, the occurrence of a twin pregnancy is an undesirable outcome in assisted reproductive technology (ART) cycles.

From the clinical point of view, the obvious alternative of repeated SET allows for similar cumulative pregnancy (P) outcomes with much lower multiple pregnancy (MP) occurrences (Kamath et al., 2020). However there are other factors that lead a proportion of patients to select DET instead of repeated SET. These include economical, psychological or even embryo quality related reasons. Consequently, there are cases where the patient will have a DET due to personal reasons, or due to the recommendation of the clinicians (lower quality of embryos). In very exceptional cases, some patients may get a triple embryo transfer (TET), 3 being the maximum number of embryos that can be transferred together according to the Spanish law. In this context, there is a need for optimizing the selection of number of embryos with the objective of finding an optimal equilibrium between high chances of a single pregnancy and minimum chances of multiple pregnancy.

As has been repeatedly discussed in this thesis, and explicitly explored in Chapter 7, the straight application of conventional ML methods returns models that do not comply with expert knowledge. In this case, we had already defined three main constraints derived from available literature:

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 (positive monotonicity). [ibid.]
2. Under any conditions MP chances cannot be higher than P chances.
3. Chances of P and MP are highly correlated with age, embryo stage, and quality (Hardarson et al., 2003; Glujovsky et al., 2016; Grøndahl et al., 2017).



To this end, IDoser (described in Chapter 4) can be applied, thus adhering to the three constraints. For the first one, IDoser calls for the assumption of monotonicity to be applicable. For the second, outcome or  $y$  will be defined as the number of gestational sacs observed around the 7th week of pregnancy, which automatically eliminates the possibility of assuming more chances of P than MP. Lastly, for the third constraint, the dosing model will be linked via a set of coefficients  $\gamma$  to the variables specified, together with more covariates ( $X$ ) that were deemed to be relevant.

## Database exploration

We used a retrospective observational database constituted by embryo transfers performed in multiple centers from Spanish private clinics. Treatments included were from both fresh and frozen embryos, resulting from both own or donated gametes, and spanned from October 2010 to September 2019. Treatments excluded were those where the embryos transferred were resultant of pre-implantational gestational tests (PGTs).

A descriptive summary of the dataset can be seen in Table 7.1, and can be categorized as:

- **Patients characteristics:** age of patient, age of partner, previous fertility (previous term and/or preterm pregnancies, miscarriages and/or life births) and presence of uterine malformations.
- **Cycle characteristics :** use of egg and/or sperm donors, cycle number, transfer number, number of correctly fertilized oocytes (2PNs) obtained, number of apt embryos obtained, state of oocytes (fresh or frozen) and age of oocytes.
- **Embryo characteristics:** embryo stage (cells or blastocyst), day of culture and quality of the embryo transferred (categorized as top/medium/low).
- **Treatment:** number of embryos transferred.
- **Outcome:** number of gestational sacs at approximately 7 weeks of gestation.

Embryo morphological quality, annotated following ASEBIR guidelines, was divided into three ordinal categories where the top quality was annotated as 1 and the low quality as 3. As there are cases where there would be two or even three embryos with their related quality value, an aggregated value of quality for all embryos transferred was needed. Given that available literature indicates that transfer of embryos of mixed quality (a good quality embryo together with low quality embryos) does not produce significative difference as compared to SET with a good quality embryo (Wintner et al., 2017; Berkhout et al., 2017; Zhu et al., 2020; Theodorou et al., 2021; Pujol et al., 2021), the value that was recorded for each case was the one corresponding to the highest ranking of the embryos transferred.

	<b>Mean±SD/Proportion(%)</b>	<b>[Min-Max]</b>
	<b>(N=8821)</b>	<b>(N=8821)</b>
<b>Age of patient (years)</b>	41,15±4,89	[22-51]
<b>Age of oocyte (years)</b>	28,12±6,57	[18-46]
<b>Age of partner (years)</b>	43,13±11,41	[18-81]
<b>Use of egg donor (%)</b>	80%	-
<b>Use of sperm donor (%)</b>	22%	-
<b>Cycle number</b>	1,45±0,84	[1-9]
<b>Transfer number</b>	1,26±0,60	[1-7]
<b>Uterine malformation (%)</b>	15%	-
<b>State of oocytes [frozen] (%)</b>	19%	-
<b>Number of previous term pregnancies</b>	0,27±0,66	[0-6]
<b>Number of previous preterm pregnancies</b>	0,01±0,09	[0-2]
<b>Number of previous miscarriages</b>	0,74±1,17	[0-9]
<b>Number of previous livebirths</b>	0,26±0,67	[0-9]
<b>Number of 2PNs</b>	5,51±2,31	[1-19]
<b>Number of apt embryos</b>	3,35±1,91	[1-17]
<b>Day of transfer</b>	3,67±1,41	[2-7]
<b>State of embyo [Blasto] (%)</b>	45%	-
<b>Embryo quality</b>	1,78±0,86	[1-3]
<b>Number of embryos transferred</b>	1,62±0,54	[1-3]
<b>Number of sacs at 7 weeks</b>	0,44±0,62	[0-3]

Table 7.1: Summary statistics of the embryo transfer database.

Finally, upon close inspection of the distribution of the outcome in the database, an unbalance was detected in favor of 0 gestational sacs observed (no pregnancy achieved). This can be seen in Figure 7.1. After a first trial with IDoser this unbalance was regarded detrimental for the learning of the model, and balancing strategies were used in the training portion of the database. Specifically, random undersampling was used in the cases with 0 gestational sacs, and random oversampling in the cases with more than 1 sac. The balancing strategies aimed to have the same number of cases for outcomes 0, 1 and 2, and only randomly oversampled the cases with 3 gestational sacs to have 5 times more than the ones originally found in the database. Final distribution of cases per outcome can be found in Figure 7.2. The portion reserved for testing the trained model was left untouched.

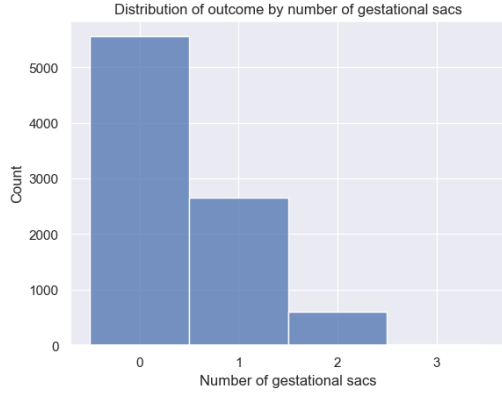


Figure 7.1: Distribution of cases by the number of gestational sacs observed around the 7th week of pregnancy in the whole database.

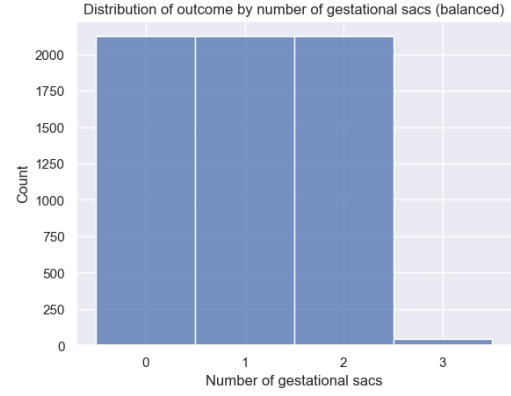


Figure 7.2: Distribution of cases by the number of gestational sacs observed around the 7th week of pregnancy after balancing strategies in the training database.

### Core model and loss function

IDoser has two main elements for its application: the core dosing model and the loss function. In the initial exploration of this problem the core model was defined as:

$$\hat{d}_i = \gamma^T x_i \quad (7.1)$$

Where for every patient  $p_i$ , the recommended dose  $\hat{d}_i$  is linearly related to the covariates  $x_i$  through a set of coefficients  $\gamma$ .

As for the loss function, the objective outcome or  $y^*$  is a single pregnancy or, in other terms, a single gestational sac observed around the 7th week of pregnancy. This means mathematically that  $y^*$  is equal to 1. Hence any outcome lower or greater than  $y^* = 1$  is considered undesirable. The allowed space for  $\hat{d}_i$  is  $[1 - 3]$  as per the laws in Spain. Any change in the number of embryos transferred (here  $d$ ) that is considered to be on the wrong direction (as per the assumed monotonicity) is not encouraged by assigning a positive loss value. Finally, to account for the uncertainty of the effect of large changes on dosification, changes bigger than 1 were not rewarded with a negative loss or  $l$  value. This translates into:

$$y_i < y^* \rightarrow \begin{cases} l = -1 & \hat{d}_i > d_i \text{ by } 1 \\ l = +1 & \hat{d}_i \neq d_i \\ l = 0 & \text{no change or increment of } d \text{ by } > 1 \end{cases} \quad (7.2)$$

$$y_i > y^* \rightarrow \begin{cases} l = -0.75 & \hat{d}_i > d_i \text{ by } 1 \\ l = +1 & \hat{d}_i \downarrow d_i \\ l = 0 & \text{no change or increment of } d \text{ by } > 1 \end{cases} \quad (7.3)$$

The  $l$  value for a reduction of  $d$  by 1 whenever  $y_i > y^*$  was slightly greater than in the reverse case to account for the barely higher number of cases where  $y_i > 1$ .

## Results

After balancing the training dataset and setting the core dosing model and loss function as detailed in the previous section, the performance of IDoserET was analyzed both by statistical comparison of *collective loss* or  $L$  defined already as:

$$L = \frac{\sum_{i=1}^N l(y_i, y_i^*, \pi(x_i), d_i)}{N}, \quad (7.4)$$

and by graphical and tabular exploration of the dose changes recommended. In the first case, results showed that the  $L$  achieved is negative. This means that improvement as compared to clinical practice (where  $l$  is always 0, and so  $L = 0$  too). But when testing the hypothesis that the value of  $L$  for IDoserET is significantly smaller than the value for clinical practice, no differences were observed (see Table 7.2).

	Clinical practice	IDoserET	p-value
$L$	0	0,02	0.82

Table 7.2:  $L$  values for both the number of embryos prescribed by clinical practice and recommended by IDoserET. Differences statistically compared with the Signed-rank Wilcoxon test for  $l$  values for each test case. A p-value under 0.05 was considered significant.

When the changes (increases and reductions in the number of embryos to be transferred) recommended by IDoserET in the test database are plotted, a clear separation between the distribution of reductions and increases is observed (see Figure 7.3), where increases are majority whenever the number of sacs is over 1, and decreases are majority whenever the number of sacs is lower than 1.

In more detail this also can be observed in Table 7.3.

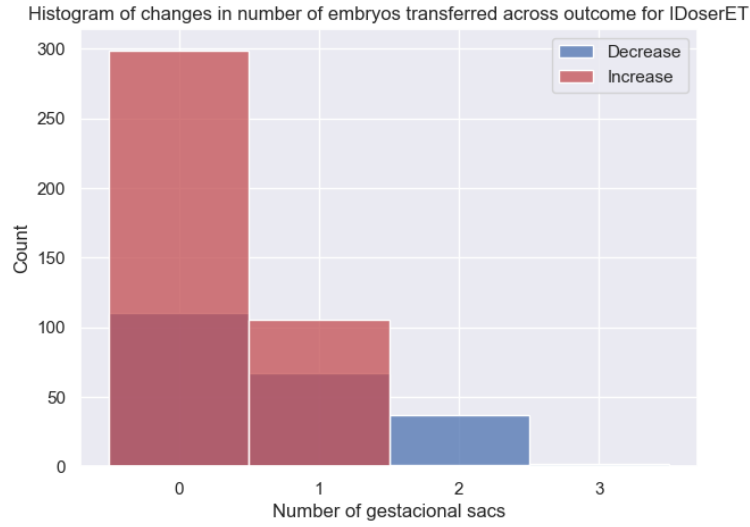


Figure 7.3: Distribution of changes on the number of embryos to be transferred recommended by IDoserET per number of gestational sacs observed.

	0 sacs	1 sac	2 sacs	3 sacs
<b>Increase 2</b>	13	0	0	0
<b>Increase 1</b>	286	105	1	0
<b>Equal</b>	702	359	83	0
<b>Decrease 1</b>	108	66	37	2
<b>Decrease 2</b>	2	1	0	0

Table 7.3: Number of embryos recommended by IDoserET as compared to clinical practice (rows) separated by the real outcome observed (columns). The cells where the changes are considered good are colored in green, in red the changes considered bad, and in yellow those where results are considered uncertain.

## Discussion

Although the  $L$  obtained in the results of the experimentation presented in the previous section indicated that an improvement was accomplished versus clinical practice, the difference was objectively small, and indeed, non significant. When the distribution of changes is analyzed in Figure 7.3 and Table 7.3, it is clear that IDoserET is capable of identifying correctly a significant portion of the cases where a reduction of the number of embryos transferred is needed (31,70%) only misidentifying a 0,81% of them. As a counterpart, it does increase the number of embryos for cases where no pregnancy was achieved in a 25,75% of cases, but decreases the number incorrectly for a 9,9% of them. Additionally, 1,17% of patients get recommended an increase in the number of embryos by 2, which could lead them either to a single or a multiple pregnancy. These changes are not considered good, as their effect is too uncertain. Finally, a 32,39% of patients that did not need any change where recommended a change by IDoser, which would be most probably detrimental for them.

In conclusion, IDoserET does identify properly which cases does need a change and recommends it in the adequate direction in a good proportion of cases, which is why a negative  $L$  value was achieved, but the number of cases where recommendations are incorrect is still too high for this model to be considered clinically acceptable. More work needs to be done in the future, with change in the core dosing model, loss function or balancing strategies as the most probable next steps. Other approaches could include the incorporation of more variables regarding the quality of the transferred embryos, like data related to their morphokinetics.

#### **7.4.4 Final conclusions and next steps**

The core objective of this thesis's work has been to understand better the underlying physiological processes like dose-response relationships in order to help every individual patients as much as possible. A step closer to the underlying truth means that we are step closer to help those patients to achieving their desire, a healthy baby at home. In this context, AI and ML methods are powerful tools that clearly can help us professionals in this endeavor, but "great power comes with great responsibility" (Stan Lee). Thus, conscious collective efforts need to be done to adhere to previously achieved knowledge in order for AI models to be of real use, and not just fancy but ultimately useless, or even worse, harmful in practice. This concept goes hand in hand with the needed investment in transparency whenever AI is used, especially for healthcare, as adherence to field knowledge should also be easily proven and interpreted.

The proposals presented here are intended to benefit the patient and to support the professional, that is why they are designed to adhere to clinical knowledge and to be carefully examined before they are released for use in clinical practice. In the next steps of this line of research, IDoserFSH needs to be tested prospectively to prove its safety and for local authorities to approve its real use in patients. Regarding the method in itself, IDoser, more work needs to be invested to close the gap with physiological reality. This could also help with its implementation in IDoserET, where the partial results presented in the previous section could be improved. Furthermore, expansion of its utility to other dose-response situations outside of a one-time selection of dose, like doses that need to be adjusted over time, can be considered.

To conclude, it is our hope that the ideas described in this thesis can be of help for other researchers in their endeavours to apply AI in a mindful way in ART or any other healthcare setting. We encourage future projects to build upon our work, addressing the gaps and challenges that remain, and extending the boundaries of knowledge in this ever-evolving field. But specially, we hope that the progress we have achieved reaches their intended beneficiaries, IVF patients. If, in any way, this project serves to help them—providing them with improved outcomes or enhanced care—then our primary objective will be genuinely fulfilled.

# **Appendices**

## Appendix A

### Tables and figures for the FSH performance score function

Recommended dose rank	Real dose rank			
	100-150 IU	175-200 IU	225-250 IU	>250 IU
100-150 IU	-1	-1	-1	-1
175-200 IU	-0.2	-0.90	-0.95	-0.95
225-250 IU	0.1	-0.01	-0.7	-0.9
>250 IU	0.15	0	0	-0.01

Table A.1: Score values for every prescribed/recommended dose rank given that the result was 0 MII

Recommended dose rank	Real dose rank			
	100-150 IU	175-200 IU	225-250 IU	>250 IU
100-150 IU	-0.8	-0.9	-0.95	-0.99
175-200 IU	-0.05	-0.60	-0.85	-0.9
225-250 IU	0.3	0	-0.5	-0.85
>250 IU	0.4	0.1	0	-0.001

Table A.2: Score values for every prescribed/recommended dose rank given that the result was 6 MII.



Recommended dose rank	Real dose rank			
	100-150 IU	175-200 IU	225-250 IU	>250 IU
100-150 IU	-0.05	-0.75	-0.9	-0.95
175-200 IU	0.1	-0.01	-0.75	-0.8
225-250 IU	0.4	0.1	-0.01	-0.7
>250 IU	0.6	0.4	0.3	0

Table A.3: Score values for every prescribed/recommended dose rank given that the result was 10 MII.

Recommended dose rank	Real dose rank			
	100-150 IU	175-200 IU	225-250 IU	>250 IU
100-150 IU	0.001	-0.2	-0.4	-0.7
175-200 IU	0.2	0.01	-0.15	-0.6
225-250 IU	0.7	0.6	0.05	-0.15
>250 IU	0.85	0.8	0.8	0.2

Table A.4: Score values for every prescribed/recommended dose rank given that the result was 15 MII.

Recommended dose rank	Real dose rank			
	100-150 IU	175-200 IU	225-250 IU	>250 IU
100-150 IU	0.01	0	0	-0.1
175-200 IU	0.8	0.40	0	-0.01
225-250 IU	0.9	0.85	0.8	0.1
>250 IU	1	1	1	1

Table A.5: Score values for every prescribed/recommended dose rank given that the result was 25 MII.

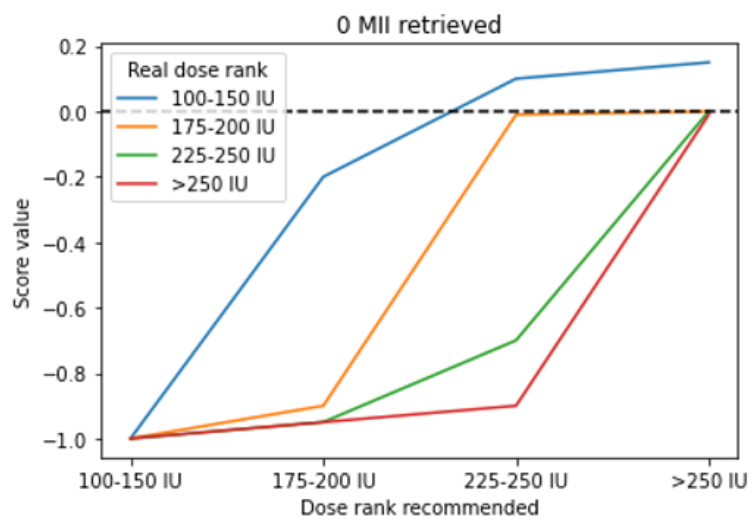


Figure A.1: Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 0 MII retrieved.

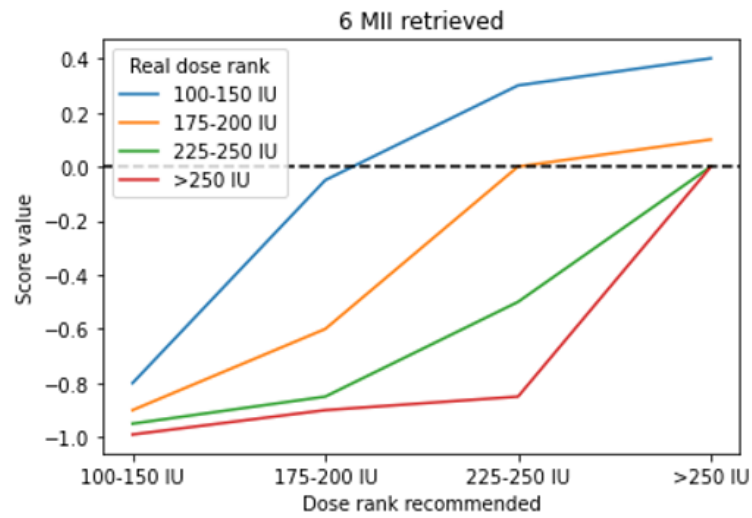


Figure A.2: Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 6 MII retrieved.

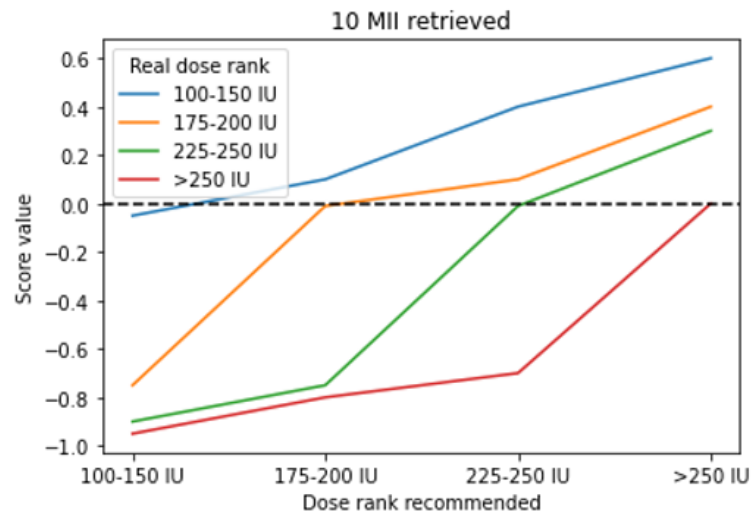


Figure A.3: Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 10 MII retrieved.

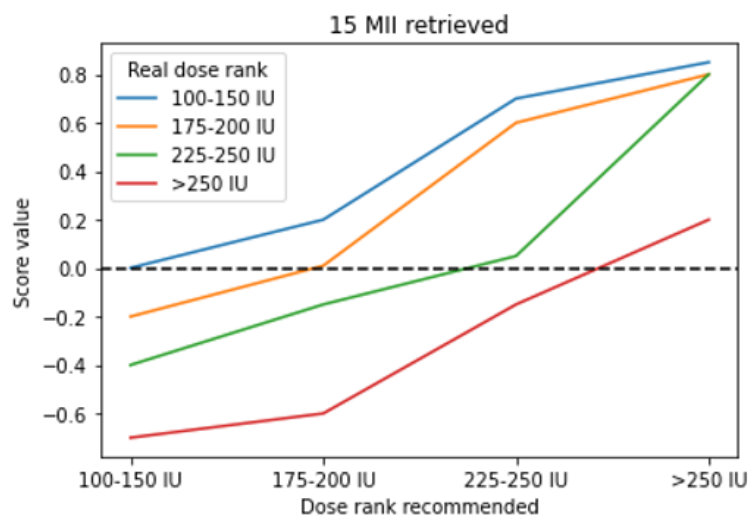


Figure A.4: Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 15 MII retrieved.

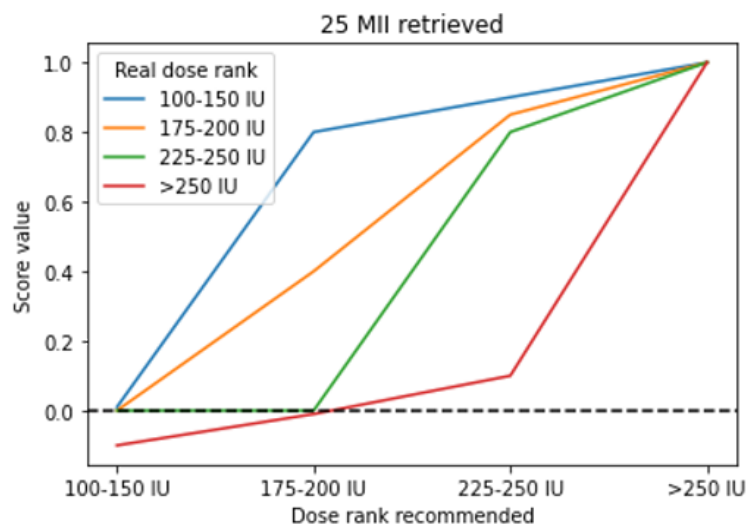


Figure A.5: Linear representation of the scores of all combinations of prescription/recommended dose ranks given that the outcome was 25 MII retrieved.

## Appendix B

# IDoser: Optimization exploration and statistic results

### B.1 Optimization exploration

Given that by our definition  $y_{min}^*$  and  $y_{max}^*$  are slightly higher (10 to 15) we primarily used their  $\beta$  coefficients and optimized only  $\kappa$ . Next, our workflow was designed to explore optimization of all coefficients in  $\gamma$  and addition/omission of covariates. Specifically:

1. Optimization of only  $\kappa$
2. Optimization of all  $\gamma$
3. Addition of AFC and BMI covariates, available in our database
4. Omission of basal FSH covariate

This resulted in 4 new optimized dosing models to be compared to the benchmark and the clinical dosing policy recorded in our database, which we will refer to from now on as *baseline*. Once  $\gamma$  values were obtained for all 4 models, a secondary optimization was run to automatically find an upper bound to dose or  $d_{max}^*$ , as a further measure for a safe and conservative model. Every model was trained with all available data depending on the covariates included, but validated always on the same database where all covariates were filled in, to avoid possible biases on the population. The benchmark and our 4 iterations were validated across 4 possible  $d_{max}$ : 300, 350, 400, and 450. For each limit, validation cases with  $d$  up to that value were admitted, and every model was allowed to dose at a maximum of the same value, and every individual loss evaluated. It is worth noting here that our models were autolimiting themselves with their optimized value of  $d_{max}^*$ , whenever this value was lower than any of 4  $d_{max}$  explored.

In the end, only the best of all 4 iterations was selected as our final proposal.

## **B.2 Statistics results**

<b>mean loss</b>	0	0.1423221	IDoser $\kappa$	IDoser all variables	IDoser + AFC + BMI	IDoser without FSH
<b>Clinical Practice</b>						
<b>La Marca</b>	0.02998277*					
<b>IDoser <math>\kappa</math></b>	0.85500958	0.043379244*				
<b>IDoser all variables</b>	0.54500150	0.004820011*	0.50871801			
<b>IDoser+AFC+BMI</b>	0.66092247	0.005758714*	0.55620182	0.818820510		
<b>IDoser without FSH</b>	0.54500150	0.004820011*	0.50871801	0.953885933	0.804125364	

Table B.1: Posthoc test of differences in individual losses between explored methods and clinical baseline capping  $d_{max}$  at 300, p-values adjusted using Finner method and marked with \* when under 0.05.

<b>mean loss</b>	0	-0.1301115	IDoser $\kappa$	IDoser all variables	IDoser + AFC + BMI	IDoser without FSH
<b>Clinical Practice</b>						
<b>La Marca</b>	3.402732e-02*					
<b>IDoser <math>\kappa</math></b>	3.243974e-06*	0.017203405*				
<b>IDoser all variables</b>	9.522397e-08*	0.001441322*	5.058417e-01			
<b>IDoser+AFC+BMI</b>	3.703982e-07*	0.004658995*	7.258543e-01	7.258543e-01		
<b>IDoser without FSH</b>	9.522397e-08*	0.001441322*	5.058417e-01	9.724247e-01	7.258543e-01	

Table B.2: Posthoc test of differences in individual losses between explored methods and clinical baseline capping  $d_{max}$  at 350, p-values adjusted using Finner method and marked with \* when under 0.05.

<b>mean loss</b>	<b>Clinical Practice</b>	<b>La Marca</b>	<b>IDoser <math>\kappa</math></b>	<b>IDoser all variables</b>	<b>IDoser + AFC + BMI</b>	<b>IDoser without FSH</b>
	0	0.07777778	-0.2444444	-0.3	-0.2925926	-0.314814
<b>Clinical Practice</b>						
<b>La Marca</b>	9.062553e-01					
<b>IDoser <math>\kappa</math></b>	1.439933e-05*	8.548517e-06*				
<b>IDoser all variables</b>	5.412448e-07*	5.412448e-07*	5.366142e-01			
<b>IDoser+AFC+BMI</b>	1.790016e-06*	1.046312e-06*	7.413683e-01	7.582353e-01		
<b>IDoser without FSH</b>	5.412448e-07*	5.412448e-07*	5.366142e-01	9.816485e-01	7.582353e-01	

Table B.3: Posthoc test of differences in individual losses between explored methods and clinical baseline capping  $d_{max}$  at 400, p-values adjusted using Finner method and marked with \* when under 0.05.

<b>mean loss</b>	<b>Clinical Practice</b>	<b>La Marca</b>	<b>IDoser <math>\kappa</math></b>	<b>IDoser all variables</b>	<b>IDoser + AFC + BMI</b>	<b>IDoser without FSH</b>
	0	0.08791209	-0.2307692	-0.2857143	-0.2783883	-0.3003663
<b>Clinical Practice</b>						
<b>La Marca</b>	7.585558e-01					
<b>IDoser <math>\kappa</math></b>	3.540466e-05*	5.550225e-06*				
<b>IDoser all variables</b>	1.245251e-06*	3.574535e-07*	5.473513e-01			
<b>IDoser+AFC+BMI</b>	5.470478e-06*	1.245251e-06*	7.585558e-01	7.585558e-01		
<b>IDoser without FSH</b>	1.245251e-06*	3.574535e-07*	5.473513e-01	9.726274e-01	7.585558e-01	

Table B.4: Posthoc test of differences in individual losses between explored methods and clinical baseline capping  $d_{max}$  at 450, p-values adjusted using Finner method and marked with \* when under 0.05.





# Bibliography

- Abbara, Ali, Aaran Patel, Tia Hunjan, Sophie A. Clarke, Germaine Chia, Pei Chia Eng, Maria Phylactou, Alexander N. Comninou, Stuart Lavery, Geoffrey H. Trew, Rehan Salim, Raj S. Rai, Tom W. Kelsey, and Waljit S. Dhillon (2019). “FSH Requirements for Follicle Growth During Controlled Ovarian Stimulation”. In: *Frontiers in Endocrinology* 10.August, pp. 1–11. DOI: 10.3389/fendo.2019.00579 (cit. on p. 31).
- Abd-Elaziz, Khalid, Ingrid Duijkers, Lars Stöckl, Bruno Dietrich, Christine Klipping, Kelvin Eckert, and Steffen Goletz (2017). “A new fully human recombinant FSH (follitropin epsilon): Two phase i randomized placebo and comparator-controlled pharmacokinetic and pharmacodynamic trials”. In: *Human Reproduction* 32.8, pp. 1639–1647. DOI: 10.1093/humrep/dex220 (cit. on pp. 31, 34, 63).
- Abdalla, H. and M. Y. Thum (2004). “An elevated basal FSH reflects a quantitative rather than qualitative decline of the ovarian reserve”. In: *Human Reproduction* 19.4, pp. 893–898. DOI: 10.1093/humrep/deh141 (cit. on p. 31).
- “Abstracts of the 34rd Annual Meeting of the European Society of Human Reproduction and Embryology” (2017). In: *Human Reproduction* 33.suppl.1, pp. i1–i541. DOI: 10.1093/humrep/33.Supplement\1.1 (cit. on p. 2).
- Aggarwal, Abhishek, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao (2022). “Artificial Intelligence (AI)-based Chatbots in Promoting Health Behavioral Changes: A Systematic Review”. In: *medRxiv* 25, p. 2022.07.05.22277263. DOI: 10.2196/40789 (cit. on p. 8).
- Ahmed, Adnan, Rishi Charate, and Naga Venkata K Pothineni (2020). “Role of Digital Health During Corona virus Disease 2019 Pandemic and Future Perspectives”. In: *Card Electrophysiol Clin* January. DOI: <https://doi.org/10.1016/j.ccep.2021.10.013> (cit. on p. 8).
- Ahmed, Arfan, Sarah Aziz, Mahmood Alzubaidi, Jens Schneider, and Sara Irshaidat (2020). “Wearable devices for anxiety & depression: A scoping review”. In: *Computer Methods and Programs in Biomedicine Update* January. DOI: <https://doi.org/10.1016/j.cmpbup.2023.100095> (cit. on p. 8).
- Aldemir, Oya, Runa Ozelci, Emre Baser, Iskender Kaplanoglu, Serdar Dilbaz, Berna Dilbaz, and Ozlem Moraloglu Tekin (2020). “Impact of Transferring a Poor Quality Embryo along with a Good Quality Embryo on Pregnancy Outcomes in IVF/ICSI Cycles: a Retrospective Study”.

- In: *Geburtshilfe und Frauenheilkunde* 80.8, pp. 844–850. DOI: 10.1055/a-1213-9164 (cit. on pp. 88, 89).
- Allegra, Adolfo, Angelo Marino, Aldo Volpes, Francesco Coffaro, Piero Scaglione, Salvatore Gullo, and Antonio La Marca (2017). “A randomized controlled trial investigating the use of a predictive nomogram for the selection of the FSH starting dose in IVF/ICSI cycles”. In: *Reproductive BioMedicine Online* 34.4, pp. 429–438. DOI: 10.1016/j.rbmo.2017.01.012 (cit. on pp. 24, 32, 37, 39, 52, 54, 65, 74).
- Amanvermez, Ramazan and Migraci Tosun (2015). *An update on ovarian aging and ovarian reserve tests* (cit. on p. 31).
- Anderson, R. A., S. M. Nelson, and W. H.B. Wallace (2012). “Measuring anti-Müllerian hormone for the assessment of ovarian reserve: When and for whom is it indicated?” In: *Maturitas* 71.1, pp. 28–33. DOI: 10.1016/j.maturitas.2011.11.008 (cit. on p. 31).
- Arce, Joan Carles, Bjarke M. Klein, and Lars Erichsen (2016). “Using amh for determining a stratified gonadotropin dosing regimen for IVF/ICSI and optimizing outcomes”. In: *Anti-Müllerian Hormone: Biology, Role in Ovarian Function and Clinical Significance*, pp. 83–102 (cit. on pp. 30, 63).
- Arce, Joan Carles, Søren Ziebe, Kersti Lundin, Ronny Janssens, Lisbeth Helmgård, and Per Sørensen (2006). “Interobserver agreement and intraobserver reproducibility of embryo quality assessments”. In: *Human Reproduction* 21.8, pp. 2141–2148. DOI: 10.1093/humrep/del106 (cit. on p. 26).
- Armstrong, D T and J H Dorrington (1976). “Androgens augment FSH-induced progesterone secretion by cultured rat granulosa cells.” eng. In: *Endocrinology* 99.5, pp. 1411–1414. DOI: 10.1210/endo-99-5-1411 (cit. on p. 30).
- Armstrong, Sarah, Priya Bhide, Vanessa Jordan, Allan Pacey, and Cindy Farquhar (2018). “Time-lapse systems for embryo incubation and assessment in assisted reproduction”. In: *Cochrane Database of Systematic Reviews* 2018.5. DOI: 10.1002/14651858.CD011320.pub3 (cit. on pp. 27, 37).
- ASRM (2021). “Practice Committee of the American Society for Reproductive Medicine and the Practice Committee for the Society for Assisted Reproductive Technology. Guidance regarding gamete and embryo donation”. In: *Fertility and sterility* 115.6 (cit. on p. 31).
- Baerwald, Angela R., Gregg P. Adams, and Roger A. Pierson (2012). “Ovarian antral folliculogenesis during the human menstrual cycle: A review”. In: *Human Reproduction Update* 18.1, pp. 73–91. DOI: 10.1093/humupd/dmr039 (cit. on p. 18).
- Balaban, Baak et al. (2011). “The Istanbul consensus workshop on embryo assessment: Proceedings of an expert meeting”. In: *Human Reproduction* 26.6, pp. 1270–1283. DOI: 10.1093/humrep/der037 (cit. on p. 26).

- Barnes, Josue et al. (2023). “A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study”. In: *The Lancet Digital Health* 5.1, e28–e40. DOI: 10.1016/S2589-7500(22)00213-8 (cit. on p. 27).
- Barrenetxea, Gorka, Cora Hernández, Julio Herrero, Luis Martínez Navarro, Manuel Muñoz, José María Rubio, Fernando Sánchez, and Jesús Zabaleta (2023). “Use of gonadotropins in ovarian stimulation in Spain: Delphi consensus”. In: *Journal of obstetrics and gynaecology : the journal of the Institute of Obstetrics and Gynaecology* 43.1, p. 2174692. DOI: 10.1080/01443615.2023.2174692 (cit. on p. 32).
- Barrett, Jeffrey S., Michael J. Fossler, K. David Cadieu, and Marc R. Gastonguay (2008). “Pharmacometrics: A multidisciplinary field to facilitate critical thinking in drug development and translational research settings”. In: *Journal of Clinical Pharmacology* 48.5, pp. 632–649. DOI: 10.1177/0091270008315318 (cit. on p. 12).
- Basile, Natalia, Maria Del Carmen Nogales, Fernando Bronet, Mireia Florensa, Marissa Riqueiros, Lorena Rodrigo, Juan García-Velasco, and Marcos Meseguer (2014). “Increasing the probability of selecting chromosomally normal embryos by time-lapse morphokinetics analysis”. In: *Fertility and Sterility* 101.3, 699–704.e1. DOI: 10.1016/j.fertnstert.2013.12.005 (cit. on pp. 24, 27).
- Bastu, Ercan, Faruk Buyru, Mehmet Ozsurmeli, Irem Demiral, Murat Dogan, and John Yeh (2016). “A randomized, single-blind, prospective trial comparing three different gonadotropin doses with or without addition of letrozole during ovulation stimulation in patients with poor ovarian response”. In: *European Journal of Obstetrics and Gynecology and Reproductive Biology* 203, pp. 30–34. DOI: 10.1016/j.ejogrb.2016.05.027 (cit. on p. 40).
- Ben-Meir, Assaf, Clara Miret Lucio, Marta Lozano, Rabi Ahmed-Odia, Semra Kahraman, Yesim Kumtepe Colakoglu, Hakan Kadir Yelke, Triantafillos Triantafillou, Emilio Gomez, Danilo Cimadomo, Adelle Yun Xin Lim, Adriana Brualla Mora, Iris Har-Vardi, Anat Sakov, and Cristina Hickman (2022). “TRANSPARENT PREDICTION OF BLASTULATION, PLOIDY AND IMPLANTATION: AN INTERNATIONAL MULTISITE VALIDATION AT SIX INDEPENDENT CLINICS”. In: *Fertility and Sterility* 118.4, Supplement. 78th Scientific Congress of the American Society for Reproductive Medicine, e117. DOI: <https://doi.org/10.1016/j.fertnstert.2022.08.347> (cit. on pp. 24, 27).
- Berkhout, R P, C G Vergouw, M van Wely, A A de Melker, R Schats, S Repping, G Hamer, S Mastenbroek, and C B Lambalk (2017). “The addition of a low-quality embryo as part of a fresh day 3 double embryo transfer does not improve ongoing pregnancy rates”. In: *Human Reproduction Open* 2017.3, pp. 1–8. DOI: 10.1093/hropen/hox020 (cit. on p. 104).
- Berntsen, Jørgen, Jens Rimestad, Jacob Theilgaard Lassen, Dang Tran, and Mikkel Fly Kragh (2021). “Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences”. In: pp. 1–18. DOI: 10.1371/journal.pone.0262661 (cit. on pp. 24, 27).

- Bica, Ioana, Ahmed M. Alaa, Craig Lambert, and Mihaela van der Schaar (2021). “From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges”. In: *Clinical Pharmacology and Therapeutics* 109.1, pp. 87–100. DOI: 10.1002/cpt.1907 (cit. on p. 16).
- Bica, Ioana and James Jordon (2020). “Estimating the Effects of Continuous-valued Interventions using Generative Adversarial Networks”. In: NeurIPS (cit. on p. 16).
- Blyth, Colin R (1972). “On Simpson ’ s Paradox and the Sure-Thing Principle”. In: *Journal of the American Statistical Association* 67.338, pp. 364–366 (cit. on p. 14).
- Borges, Nuno Costa et al. (2022). “FIRST PRE-CLINICAL AND CLINICAL VALIDATION OF AN AUTOMATED SPERM INJECTION ROBOT (ICSI-A) IN HUMAN OOCYTES”. In: *Fertility and Sterility* 118.4, Supplement. 78th Scientific Congress of the American Society for Reproductive Medicine, e5. DOI: <https://doi.org/10.1016/j.fertnstert.2022.08.031> (cit. on pp. 24, 29).
- Bormann, Charles L., Manoj Kumar Kanakasabapathy, Prudhvi Thirumalaraju, Raghav Gupta, Rohan Pooniwalla, Hemanth Kandula, Eduardo Hariton, Irene Souter, Irene Dimitriadis, Leslie B. Ramirez, Carol L. Curchoe, Jason Swain, Lynn M. Boehnlein, and Hadi Shafiee (2020). “Performance of a deep learning based neural network in the selection of human blastocysts for implantation”. In: *eLife* 9, pp. 1–14. DOI: 10.7554/ELIFE.55301 (cit. on pp. 24, 27).
- Brennan, Hannah L. and Simon D. Kirby (2023). “The role of artificial intelligence in the treatment of obstructive sleep apnea”. In: *Journal of otolaryngology - head & neck surgery = Le Journal d’oto-rhino-laryngologie et de chirurgie cervico-faciale* 52.1, p. 7. DOI: 10.1186/s40463-023-00621-0 (cit. on p. 8).
- Cai, Hengrui, Chengchun Shi, Rui Song, and Wenbin Lu (2020). “Deep Jump Q-Evaluation for Offline Policy Evaluation in Continuous Action Space”. In: (cit. on p. 16).
- Calabrese, Edward J. (2016). “The emergence of the dose-response concept in biology and medicine”. In: *International Journal of Molecular Sciences* 17.12. DOI: 10.3390/ijms17122034 (cit. on p. 11).
- Calabrese, Edward J. and L. A. Baldwin (2002). “Defining hormesis”. In: *Human and Experimental Toxicology* 21.2, pp. 91–97. DOI: 10.1191/0960327102ht217oa (cit. on p. 11).
- Campbell, Alison, Simon Fishel, Natalie Bowman, Samantha Duffy, Mark Sedler, and Cristina Fontes Lindemann Hickman (2013). “Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics”. In: *Reproductive BioMedicine Online* 26.5, pp. 477–485. DOI: 10.1016/j.rbmo.2013.02.006 (cit. on pp. 24, 27).
- Cao, Qiang, Stephen Shaoyi Liao, Xiangqian Meng, Han Ye, Zhenbin Yan, and Puxi Wang (2018). “Identification of viable embryos using deep learning for medical image”. English. In: *Proceedings of 2018 5th International Conference on Bioinformatics Research and Applications*. ACM International Conference Proceeding Series. 5th International Conference on Bioinfor-

- matics Research and Applications, ICBRA 2018 ; Conference date: 27-12-2018 Through 29-12-2018. ACM New York, pp. 69–72. DOI: 10.1145/3309129.3309143 (cit. on pp. 24, 27).
- Carson, Daniel D., Indrani Bagchi, Sudhandsu K. Dey, Allen C. Enders, Asgerally T. Fazleabas, Bruce A. Lessey, and Koji Yoshinaga (2000). “Embryo implantation”. In: *Developmental Biology* 223.2, pp. 217–237. DOI: 10.1006/dbio.2000.9767 (cit. on p. 28).
- Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad (2015). “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2015-August, pp. 1721–1730. DOI: 10.1145/2783258.2788613 (cit. on p. 9).
- Chamayou, Sandrine, Maria Sicali, Carmelita Alecci, Carmen Ragolia, Annalisa Liprino, Daniela Nibali, Giorgia Storaci, Antonietta Cardea, and Antonino Guglielmino (2017). “The accumulation of vitrified oocytes is a strategy to increase the number of euploid available blastocysts for transfer after preimplantation genetic testing”. In: *Journal of Assisted Reproduction and Genetics* 34.4, pp. 479–486. DOI: 10.1007/s10815-016-0868-0 (cit. on p. 32).
- Chan, An Wen, Jennifer M. Tetzlaff, Douglas G. Altman, et al. (2013). “Spirit 2013 statement: Defining standard protocol items for clinical trials”. In: *Chinese Journal of Evidence-Based Medicine* 13.12, pp. 1501–1507. DOI: 10.7507/1672-2531.20130256 (cit. on p. 74).
- Chan, An Wen, Jennifer M. Tetzlaff, Peter C. Gøtzsche, Douglas G. Altman, Howard Mann, Jesse A. Berlin, Kay Dickersin, Asbjørn Hróbjartsson, Kenneth F. Schulz, Wendy R. Parulekar, Karmela Krleza-Jeric, Andreas Laupacis, and David Moher (2013). “SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials.” In: *BMJ (Clinical research ed.)* 346, pp. 1–42. DOI: 10.1136/bmj.e7586 (cit. on p. 74).
- Chen, Yuan hui, Qian Wang, Ya nan Zhang, Xiao Han, Dong han Li, and Cui lian Zhang (2017). “Cumulative live birth and surplus embryo incidence after frozen-thaw cycles in PCOS: how many oocytes do we need?” In: *Journal of Assisted Reproduction and Genetics* 34.9, pp. 1153–1159. DOI: 10.1007/s10815-017-0959-6 (cit. on pp. 25, 63).
- Choi, Bokyoung, Ernesto Bosch, Benjamin M. Lannon, Marie Claude Leveille, Wing H. Wong, Arthur Leader, Antonio Pellicer, Alan S. Penzias, and Mylene W.M. Yao (2013). “Personalized prediction of first-cycle in vitro fertilization success”. In: *Fertility and Sterility* 99.7, pp. 1905–1911. DOI: 10.1016/j.fertnstert.2013.02.016 (cit. on pp. 24, 29).
- Colangelo, Kyle and Ying-Ying Lee (2020). “Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments”. In: pp. 1–39 (cit. on p. 16).
- Commission, European (2022). *Ethics Guidelines for Trustworthy AI*. URL: <https://digital-st%20ategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (cit. on p. 17).
- Conaghan, Joe, Alice A. Chen, Susan P. Willman, Kristen Ivani, Philip E. Chenette, Robert Boostanfar, Valerie L. Baker, G. David Adamson, Mary E. Abusief, Marina Gvakharia, Kevin E. Loewke, and Shehua Shen (2013). “Improving embryo selection using a computer-automated

- time-lapse image analysis test plus day 3 morphology: Results from a prospective multicenter trial”. In: *Fertility and Sterility* 100.2, 412–419.e5. DOI: 10.1016/j.fertnstert.2013.04.021 (cit. on pp. 24, 27).
- Conforti, Alessandro, Sandro C. Esteves, Silvia Picarelli, Giuseppe Iorio, Erika Rania, Fulvio Zullo, Giuseppe De Placido, and Carlo Alviggi (2019). “Novel approaches for diagnosis and management of low prognosis patients in assisted reproductive technology: The POSEIDON concept”. In: *Panminerva Medica* March, pp. 24–29. DOI: 10.23736/S0031-0808.18.03511-5 (cit. on p. 31).
- Cooper, Gregory F., Vijoy Abraham, Constantin F. Aliferis, John M. Aronis, Bruce G. Buchanan, Richard Caruana, Michael J. Fine, Janine E. Janosky, Gary Livingston, Tom Mitchell, Stefano Monti, and Peter Spirtes (2005). “Predicting dire outcomes of patients with community acquired pneumonia”. In: *Journal of Biomedical Informatics* 38.5, pp. 347–366. DOI: 10.1016/j.jbi.2005.02.005 (cit. on p. 9).
- Cooper, Gregory F., Constantin F. Aliferis, Richard Ambrosino, John Aronis, Bruce G. Buchanan, Richard Caruana, Michael J. Fine, Clark Glymour, Geoffrey Gordon, Barbara H. Hanusa, Janine E. Janosky, Christopher Meek, Tom Mitchell, Thomas Richardson, and Peter Spirtes (1997). “An evaluation of machine-learning methods for predicting pneumonia mortality”. In: *Artificial Intelligence in Medicine* 9.2, pp. 107–138. DOI: 10.1016/S0933-3657(96)00367-3 (cit. on p. 9).
- Corani, G, C Magli, A Giusti, L Gianaroli, and L M Gambardella (2013). “A Bayesian network model for predicting pregnancy after in vitro fertilization.” eng. In: *Comput Biol Med* 43.11, pp. 1783–1792. DOI: 10.1016/j.compbimed.2013.07.035 (cit. on pp. 24, 27).
- Correa, Núria, Jesús Cerquides, Josep Lluís Arcos, and Rita Vassena (2022). “Supporting first FSH dosage for ovarian stimulation with machine learning”. In: *Reproductive BioMedicine Online* 45.5, pp. 1039–1045. DOI: 10.1016/j.rbmo.2022.06.010 (cit. on pp. 38, 100).
- (2023). “EP-226 Aide à la sélection de la dose de FSH pour la stimulation ovarienne à l’aide du Machine Learning”. In: Oral communication. Pari(s) Santé Femmes. Lille (cit. on p. 101).
- Correa, Núria, Jesús Cerquides, Josep Lluís Arcos, Rita Vassena, and Mina Popovic (2023a). “O-185 A clinically robust machine learning model for selecting the first FSH dose during controlled ovarian hyperstimulation: incorporating clinical knowledge to the learning process.” In: Oral communication. European Society of Human Reproduction and Embryology (ESHRE) Annual Meeting. Copenhagen (cit. on pp. 56, 101).
- (2023b). “Personalizing the first dose of FSH for IVF patients through machine learning: a non-inferiority study protocol for a multi-center randomized controlled trial”. Under review by the *Trials* journal (cit. on p. 101).
- Correa, Núria, Jesús Cerquides, Amelia Rodríguez-Aranda, Josep Lluís Arcos, and Rita Vassena (2022). “379/427 Acompañamiento en la selección de la dosis de FSH para estimulación ovárica mediante machine learning.” In: *33º Congreso Nacional Sociedad Española de Fer-*

- tilidad*. Oral communication. Sociedad Española de Fertilidad (SEF). Bilbao (cit. on pp. 38, 100).
- Correa, Núria, Jesús Cerquides, Rita Vassena, Mina Popovic, and Josep Lluís Arcos (2023). “IDoser: Improving individualized dosing policies with clinical practice and machine learning”. In: *medRxiv*. DOI: 10.1101/2023.03.28.23287859 (cit. on pp. 56, 80, 101).
- Correa, Núria, Flavia Rodríguez, Jesús Cerquides, Josep Lluís Arcos, and Rita Vassena (2021). “P-637 Development and validation of an Artificial Intelligence algorithm that matches a clinician ability to select the best follitropin dose for ovarian stimulation”. In: *Human Reproduction* 36.Supplement\_1. deab130.636. DOI: 10.1093/humrep/deab130.636 (cit. on pp. 38, 100).
- Correa, Núria, Rita Vassena, Jesús Cerquides, and Josep Lluís Arcos (2021). “Limits of Conventional Machine Learning Methods to Predict Pregnancy and Multiple Pregnancy After Embryo Transfer”. In: CCIA. DOI: 10.3233/faia210141 (cit. on pp. 86, 100).
- Crosignani, P. G. et al. (2000). “Multiple gestation pregnancy”. In: *Human Reproduction* 15.8, pp. 1856–1864. DOI: 10.1093/humrep/15.8.1856 (cit. on pp. 28, 34, 87, 103).
- Cruz Rivera, Samantha, Xiaoxuan Liu, Sarah E. Hughes, Helen Dunster, Elaine Manna, Alastair K. Denniston, and Melanie J. Calvert (2023). “Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies”. In: *The Lancet. Digital health* 5.3, e168–e173. DOI: 10.1016/S2589-7500(22)00252-7 (cit. on p. 8).
- Cuevas Saiz, Irene, Maria Carme Pons Gatell, Muriel Cuadros Vargas, Arantzazu Delgado Mendive, Natalia Rives Enedáguila, Marta Moragas Solanes, Beatriz Carrasco Canal, José Teruel López, Ana Busquets Bonet, and M<sup>a</sup> Victoria Hurtado de Mendoza Acosta (2018). “The Embryology Interest Group: updating ASEBIR’s morphological scoring system for early embryos, morulae and blastocysts”. In: *Medicina Reproductiva y Embriología Clínica* 5.1, pp. 42–54. DOI: 10.1016/j.medre.2017.11.002 (cit. on p. 27).
- Curchoe, Carol Lynn (2023). “Proceedings of the first world conference on AI in fertility”. In: *Journal of Assisted Reproduction and Genetics*, pp. 215–222. DOI: 10.1007/s10815-022-02704-9 (cit. on pp. 23, 29, 36).
- Curchoe, Carol Lynn and Charles L. Bormann (2019). “Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018”. In: *Journal of Assisted Reproduction and Genetics* 36.4, pp. 591–600. DOI: 10.1007/s10815-019-01408-x (cit. on p. 22).
- Curchoe, Carol Lynn, Jonas Malmsten, et al. (2020). “Predictive modeling in reproductive medicine: Where will the future of artificial intelligence research take us?” In: *Fertility and Sterility* 114.5, pp. 934–940. DOI: 10.1016/j.fertnstert.2020.10.040 (cit. on p. 23).
- Danardono, Gunawan B., Alva Erwin, James Purnama, Nining Handayani, Arie A. Polim, Arief Boediono, and Ivan Sini (2022). “A Homogeneous Ensemble of Robust Pre-defined Neural Network Enables Automated Annotation of Human Embryo Morphokinetics”. In: *Journal of*

- Reproduction and Infertility* 23.4, pp. 250–256. DOI: 10.18502/jri.v23i4.10809 (cit. on pp. 24, 29).
- Darwich, Adam S, Kayode Ogungbenro, Alexander A Vinks, J Robert Powell, Niloufar Marsousi, Youssef Daali, David Fairman, Jack Cook, and Lawrence J Lesko (2017). “Why has model-informed precision dosing not yet become common clinical reality ? Lessons from the past and a roadmap for the future This article has been accepted for publication and undergone full peer review but has not been through the copyediting , ty”. In: *Clin Pharmacol Ther* 101, pp. 646–656 (cit. on pp. 12, 13).
- Das, Kaushik Pratim and Chandra J (2023). “Nanoparticles and convergence of artificial intelligence for targeted drug delivery for cancer therapy: Current progress and challenges”. In: *Frontiers in Medical Technology* 4.January, pp. 1–14. DOI: 10.3389/fmedt.2022.1067144 (cit. on p. 8).
- De Geyter, Christian et al. (2018). “ART in Europe, 2014: Results generated from European registries by ESHRE”. In: *Human Reproduction* 33.9, pp. 1586–1601. DOI: 10.1093/humrep/dey242 (cit. on pp. 21, 87).
- De Gheselle, Stefanie, Céline Jacques, Jérôme Chambost, Celine Blank, Klaas Declerck, Ilse De Croo, Cristina Hickman, and Kelly Tilleman (2022). “Machine learning for prediction of euploidy in human embryos: in search of the best-performing model and predictive features”. In: *Fertility and Sterility* 117.4, pp. 738–746. DOI: 10.1016/j.fertnstert.2021.11.029 (cit. on pp. 24, 27).
- Demšar, Janez (2006). “Statistical comparisons of classifiers over multiple data sets”. In: *Journal of Machine Learning Research* 7, pp. 1–30 (cit. on p. 66).
- Denysyuk, Hanna Vitaliyivna, Rui João Pinto, Pedro Miguel Silva, Rui Pedro Duarte, Francisco Alexandre Marinho, Luís Pimenta, António Jorge Gouveia, Norberto Jorge Gonçalves, Paulo Jorge Coelho, Eftim Zdravevski, Petre Lameski, Valderi Leithardt, Nuno M. Garcia, and Ivan Miguel Pires (2023). “Algorithms for automated diagnosis of cardiovascular diseases based on ECG data: A comprehensive systematic review”. In: *Heliyon* 9.2. DOI: 10.1016/j.heliyon.2023.e13601 (cit. on p. 7).
- Diakiw, S. M., J. M.M. Hall, M. D. VerMilyea, J. Amin, J. Aizpurua, L. Giardini, Y. G. Briones, A. Y.X. Lim, M. A. Dakka, T. V. Nguyen, D. Perugini, and M. Perugini (2022). “Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF”. In: *Human Reproduction* 37.8, pp. 1746–1759. DOI: 10.1093/humrep/deac131 (cit. on pp. 24, 27).
- Diaz-Gimeno, P., P. Sebastian-Leon, J. M. Sanchez-Reyes, K. Spath, A. Aleman, C. Vidal, A. Devesa-Peiro, E. Labarta, I. Sánchez-Ribas, M. Ferrando, G. Kohls, J. A. García-Velasco, E. Seli, D. Wells, and A. Pellicer (2022). “Identifying and optimizing human endometrial gene expression signatures for endometrial dating”. In: *Human Reproduction* 37.2, pp. 284–296. DOI: 10.1093/humrep/deab262 (cit. on pp. 24, 28).



- Drakopoulos, Panagiotis, Christophe Blockeel, Dominic Stoop, Michel Camus, Michel De Vos, Herman Tournaye, and Nikolaos P. Polyzos (2016). “Conventional ovarian stimulation and single embryo transfer for IVF/ICSI. How many oocytes do we need to maximize cumulative live birth rates after utilization of all fresh and frozen embryos?” In: *Human Reproduction* 31.2, pp. 370–376. DOI: 10.1093/humrep/dev316 (cit. on p. 25).
- Ebid, Abdel Hameed I.M., Sara M. Abdel Motaleb, Mahmoud I. Mostafa, and Mahmoud M.A. Soliman (2021). “Novel nomogram-based integrated gonadotropin therapy individualization in in vitro fertilization/ intracytoplasmic sperm injection: A modeling approach”. In: *Clinical and Experimental Reproductive Medicine* 48.2, pp. 163–173. DOI: 10.5653/term.2020.03909 (cit. on pp. 24, 32, 39).
- Erlich, I., A. Ben-Meir, I. Har-Vardi, J. Grifo, F. Wang, C. Mccaffrey, D. McCulloh, Y. Or, and L. Wolf (2022). “Pseudo contrastive labeling for predicting IVF embryo developmental potential”. In: *Scientific Reports* 12.1, pp. 1–13. DOI: 10.1038/s41598-022-06336-y (cit. on pp. 24, 27).
- Eskofier, Bjoern M and Jochen Klucken (2023). “Predictive Models for Health Deterioration: Understanding Disease Pathways for Personalized Medicine”. In: pp. 131–156 (cit. on p. 8).
- Esteves, Sandro C., José F. Carvalho, Fabiola C. Bento, and Jonathan Santos (2019). “A novel predictive model to estimate the number of mature oocytes required for obtaining at least one euploid blastocyst for transfer in couples undergoing in vitro fertilization/intracytoplasmic sperm injection: The ART calculator”. In: *Frontiers in Endocrinology* 10.FEB, pp. 1–14. DOI: 10.3389/fendo.2019.00099 (cit. on pp. 24, 25, 30).
- Esteves, Sandro C., Matheus Roque, Giuliano M. Bedoschi, Alessandro Conforti, Peter Humaidan, and Carlo Alviggi (2018). *Defining low prognosis patients undergoing assisted reproductive technology: POSEIDON criteria-the why*. DOI: 10.3389/fendo.2018.00461 (cit. on pp. 47, 80).
- Fanton, Michael, Veronica Nutting, Arielle Rothman, Paxton Maeder-York, Eduardo Hariton, Oleksii Barash, Louis Weckstein, Denny Sakkas, Alan B. Copperman, and Kevin Loewke (2022). “An interpretable machine learning model for individualized gonadotrophin starting dose selection during ovarian stimulation”. In: *Reproductive BioMedicine Online* 45.6, pp. 1152–1159. DOI: 10.1016/j.rbmo.2022.07.010 (cit. on pp. 24, 33).
- Fanton, Michael, Veronica Nutting, Funmi Solano, Paxton Maeder-York, Eduardo Hariton, Oleksii Barash, Louis Weckstein, Denny Sakkas, Alan B. Copperman, and Kevin Loewke (2022). “An interpretable machine learning model for predicting the optimal day of trigger during ovarian stimulation”. In: *Fertility and Sterility* 118.1, pp. 101–108. DOI: 10.1016/j.fertnstert.2022.04.003 (cit. on pp. 24, 26).
- Fernandez, Eleonora Inácio, André Satoshi Ferreira, Matheus Henrique Miquelão Cecílio, Dóris Spinosa Chéles, Rebeca Colauto Milanezi de Souza, Marcelo Fábio Gouveia Nogueira, and José Celso Rocha (2020). “Artificial intelligence in the IVF laboratory: overview through the

- application of different types of algorithms for the classification of reproductive data”. In: *Journal of Assisted Reproduction and Genetics* 37.10, pp. 2359–2376. DOI: 10.1007/s10815-020-01881-9 (cit. on p. 23).
- Feyeux, M., A. Reignier, M. Mocaer, J. Lammers, D. Meistermann, P. Barrière, P. Paul-Gilloteaux, L. David, and T. Fréour (2020). “Development of automated annotation software for human embryo morphokinetics”. In: *Human Reproduction* 35.3, pp. 557–564. DOI: 10.1093/humrep/deaa001 (cit. on pp. 24, 29).
- Fleming, R., N. Deshpande, I. Traynor, and R. W.S. Yates (2006). “Dynamics of FSH-induced follicular growth in subfertile women: Relationship with age, insulin resistance, oocyte yield and anti-Mullerian hormone”. In: *Human Reproduction* 21.6, pp. 1436–1441. DOI: 10.1093/humrep/dei499 (cit. on p. 25).
- Fraser, Ian S, Hilary O D Critchley, Michael Broder, and Malcolm G Munro (2011). “The FIGO recommendations on terminologies and definitions for normal and abnormal uterine bleeding.” eng. In: *Semin Reprod Med* 29.5, pp. 383–390. DOI: 10.1055/s-0031-1287662 (cit. on p. 18).
- Fritz, M.A. and L. Speroff (2011). *Clinical Gynecologic Endocrinology and Infertility. 8th Edition*. Philadelphia: Lippincott Williams & Wilkins (cit. on pp. 18, 19).
- Fukunaga, Noritaka, Sho Sanami, Hiroya Kitasaka, Yuji Tsuzuki, Hiroyuki Watanabe, Yuta Kida, Seiji Takeda, and Yoshimasa Asada (2020). “Development of an automated two pronuclei detection system on time-lapse embryo images using deep learning techniques”. In: *Reproductive Medicine and Biology* 19.3, pp. 286–294. DOI: 10.1002/rmb2.12331 (cit. on pp. 24, 29).
- Gadagkar, Sudhindra R. and Gerald B. Call (2015). “Computational tools for fitting the Hill equation to dose-response curves”. In: *Journal of Pharmacological and Toxicological Methods* 71, pp. 68–76. DOI: 10.1016/j.vascn.2014.08.006 (cit. on p. 12).
- García, Salvador, Alberto Fernández, Julián Luengo, and Francisco Herrera (2010). “Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power”. In: *Information Sciences* 180.10, pp. 2044–2064. DOI: 10.1016/j.ins.2009.12.010 (cit. on p. 66).
- García, Salvador and Francisco Herrera (2008). “An extension on ”statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons”. In: *Journal of Machine Learning Research* 9.May, pp. 2677–2694 (cit. on p. 66).
- Gardner, David K., Michelle Lane, John Stevens, Terry Schlenker, and William B. Schoolcraft (2000). “Blastocyst score affects implantation and pregnancy outcome: Towards a single blastocyst transfer”. In: *Fertility and Sterility* 73.6, pp. 1155–1158. DOI: 10.1016/S0015-0282(00)00518-5 (cit. on p. 26).
- Gautam, Nitesh et al. (2022). “Artificial Intelligence, Wearables and Remote Monitoring for Heart Failure: Current and Future Applications”. In: *Diagnostics* 12.12, pp. 1–19. DOI: 10.3390/diagnostics12122964 (cit. on p. 8).

- Geampana, Alina and Manuela Perrotta (2023). “Predicting Success in the Embryology Lab: The Use of Algorithmic Technologies in Knowledge Production”. In: *Science Technology and Human Values* 48.1, pp. 212–233. DOI: 10.1177/01622439211057105 (cit. on p. 23).
- Gianaroli, L., M. C. Magli, L. Gambardella, A. Giusti, C. Grugnetti, and G. Corani (2013). “Objective way to support embryo transfer: A probabilistic decision”. In: *Human Reproduction* 28.5, pp. 1210–1220. DOI: 10.1093/humrep/det030 (cit. on pp. 24, 27).
- Glujovsky, Demián, Cindy Farquhar, Andrea Marta Quinteiro Retamar, Cristian Roberto Alvarez Sedo, and Deborah Blake (2016). “Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology”. In: *Cochrane Database of Systematic Reviews* 2016.6. DOI: 10.1002/14651858.CD002118.pub5 (cit. on pp. 34, 88, 103).
- Gonzalez, Daniel, Gauri G. Rao, Stacy C. Bailey, Kim L.R. Brouwer, Yanguang Cao, Daniel J. Crona, Angela D.M. Kashuba, Craig R. Lee, Kathryn Morbitzer, J. Herbert Patterson, Tim Wiltshire, Jon Easter, Scott W. Savage, and J. Robert Powell (2017). “Precision Dosing: Public Health Need, Proposed Framework, and Anticipated Impact”. In: *Clinical and Translational Science* 10.6, pp. 443–454. DOI: 10.1111/cts.12490 (cit. on pp. 13, 57).
- Grøndahl, Marie Louise, Sofie Lindgren Christiansen, Ulrik Schiøler Kesmodel, Inge Errebo Agerholm, Josephine Gabriela Lemmen, Peter Lundstrøm, Jeanette Bogstad, Morten Raaschou-Jensen, and Steen Ladelund (2017). “Effect of women’s age on embryo morphology, cleavage rate and competence - A multicenter cohort study”. In: *PLoS ONE* 12.4, pp. 1–12. DOI: 10.1371/journal.pone.0172456 (cit. on pp. 34, 88, 103).
- Gui, Yu, Xiujing He, Jing Yu, and Jing Jing (2023). “Artificial Intelligence-Assisted Transcriptomic Analysis to Advance Cancer Immunotherapy”. In: *Journal of Clinical Medicine* 12.4. DOI: 10.3390/jcm12041279 (cit. on p. 8).
- Hamberg, Anna Karin, Jacob Hellman, Jonny Dahlberg, E. Niclas Jonsson, and Mia Wadelius (2015). “A Bayesian decision support tool for efficient dose individualization of warfarin in adults and children”. In: *BMC Medical Informatics and Decision Making* 15.1, pp. 1–9. DOI: 10.1186/s12911-014-0128-0 (cit. on p. 13).
- Hansen, Karl R., George M. Hodnett, Nicholas Knowlton, and Latasha B. Craig (2011). “Correlation of ovarian reserve tests with histologically determined primordial follicle number”. In: *Fertility and Sterility* 95.1, pp. 170–175. DOI: 10.1016/j.fertnstert.2010.04.006 (cit. on p. 31).
- Haouzi, Delphine, Hervé Dechaud, Said Assou, John De Vos, and Samir Hamamah (2012). “Insights into human endometrial receptivity from transcriptomic and proteomic data”. In: *Reproductive BioMedicine Online* 24.1, pp. 23–34. DOI: 10.1016/j.rbmo.2011.09.009 (cit. on p. 28).
- Hardarson, Thorir, Gunilla Caisander, Anita Sjögren, Charles Hanson, Lars Hamberger, and Kersti Lundin (2003). “A morphological and chromosomal study of blastocysts developing from morphologically suboptimal human pre-embryos compared with control blastocysts”. In: *Human Reproduction* 18.2, pp. 399–407. DOI: 10.1093/humrep/deg092 (cit. on pp. 34, 88, 103).

- Hariton, Eduardo, Ethan A. Chi, Gordon Chi, Jerrine R. Morris, Jon Braatz, Pranav Rajpurkar, and Mitchell Rosen (2021). “A machine learning algorithm can optimize the day of trigger to improve in vitro fertilization outcomes”. In: *Fertility and Sterility* 116.5, pp. 1227–1235. DOI: 10.1016/j.fertnstert.2021.06.018 (cit. on pp. 24, 26).
- Harper, M J (1992). “The implantation window.” eng. In: *Baillieres Clin Obstet Gynaecol* 6.2, pp. 351–371. DOI: 10.1016/s0950-3552(05)80092-6 (cit. on p. 28).
- Harrison, Robert F., Saji Jacob, Helen Spillane, Eimear Mallon, and Bernadette Hennelly (2001). “A prospective randomized clinical trial of differing starter doses of recombinant follicle-stimulating hormone (follitropin- $\beta$ ) for first time in vitro fertilization and intracytoplasmic sperm injection treatment cycles”. In: *Fertility and Sterility* 75.1, pp. 23–31. DOI: 10.1016/S0015-0282(00)01643-5 (cit. on p. 40).
- Hayes, A. Wallace, Tao Wang, and Darlene Dixon (2020). “Chapter 2 - Dose and dose-response relationships in toxicology”. In: *Loomis’s Essentials of Toxicology (Fifth Edition)*. Ed. by A. Wallace Hayes, Tao Wang, and Darlene Dixon. Fifth Edition. Academic Press, pp. 17–31. DOI: <https://doi.org/10.1016/B978-0-12-815921-7.00002-8> (cit. on p. 11).
- He, Aihua et al. (2023). “Can biomarkers identified from the uterine fluid transcriptome be used to establish a noninvasive endometrial receptivity prediction tool? A proof-of-concept study”. In: *Reproductive Biology and Endocrinology* 21.1, pp. 1–15. DOI: 10.1186/s12958-023-01070-0 (cit. on pp. 24, 28).
- Hernández-González, Jerónimo, Iñaki Inza, Lorena Crisol-Ortíz, María A Guembe, María J. Iñarra, and Jose A. Lozano (2018). “Fitting the data from embryo implantation prediction: Learning from label proportions.” eng. In: *Stat Methods Med Res* 27.4, pp. 1056–1066. DOI: 10.1177/0962280216651098 (cit. on pp. 24, 27).
- Hernández-González, Jerónimo, Olga Valls, Adrián Torres-Martín, and Jesús Cerquides (2022). “Modeling three sources of uncertainty in assisted reproductive technologies with probabilistic graphical models”. In: *Computers in Biology and Medicine* 150.April, p. 106160. DOI: 10.1016/j.combiomed.2022.106160 (cit. on pp. 24, 27).
- Hickman, Cristina, Nikica Zaninovic, Jonas Malmsten, Qiansheng Zhan, Adriana Brualla Mora, Iris Har-Vardi, and Assaf Ben-Meir (2022). “TURNING THE BLACK BOX INTO A GLASS BOX: USE OF TRANSPARENT ARTIFICIAL INTELLIGENCE TO UNDERSTAND BIOLOGICAL MARKERS USEFUL FOR EMBRYO SELECTION”. In: *Fertility and Sterility* 118.4, Supplement. 78th Scientific Congress of the American Society for Reproductive Medicine, e5–e6. DOI: <https://doi.org/10.1016/j.fertnstert.2022.08.032> (cit. on p. 27).
- Hillier, S G, L E Jr Reichert, and E V Van Hall (1981). “Control of preovulatory follicular estrogen biosynthesis in the human ovary.” eng. In: *J Clin Endocrinol Metab* 52.5, pp. 847–856. DOI: 10.1210/jcem-52-5-847 (cit. on p. 18).
- Hirano, Keisuke and Guido W. Imbens (2005). “The Propensity Score with Continuous Treatments”. In: *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspec-*

- tives: An Essential Journey with Donald Rubin's Statistical Family* 2001, pp. 73–84. DOI: 10.1002/0470090456.ch7 (cit. on p. 16).
- Hosseini, Mohsen Masoumian, Seyedeh Toktam, Masoumian Hosseini, Karim Qayumi, and Soleiman Ahmady (2023). “The Aspects of Running Artificial Intelligence in Emergency Care ; a Scoping Review”. In: 11.1, pp. 1–28 (cit. on p. 8).
- Howles, C. M., H. Saunders, V. Alam, and P. Engrand (2006). “Predictive factors and a corresponding treatment algorithm for controlled ovarian stimulation in patients treated with recombinant human follicle stimulating hormone (follitropin alfa) during assisted reproduction technology (ART) procedures. An analysis ”. In: *Current Medical Research and Opinion* 22.5, pp. 907–918. DOI: 10.1185/030079906X104678 (cit. on pp. 24, 32, 37, 39, 52).
- Iman, Ronald L. and James M. Davenport (1980). “Approximations of the critical region of the friedman statistic”. In: *Communications in Statistics - Theory and Methods* 9.6, pp. 571–595. DOI: 10.1080/03610928008827904 (cit. on p. 66).
- Imbens, G.W. (2000). “The role of propensity score in estimating dose-response functions”. In: *Biometrika* 87, pp. 706–710 (cit. on p. 16).
- Interat, Majdi, Ashok Agarwal, Sandro C. Esteves, Jenna Meyer, and Avi Harlev (2019). “Impact of Body Mass Index on female fertility and ART outcomes”. In: *Panminerva Medica* 61.1. DOI: 10.23736/s0031-0808.18.03490-0 (cit. on p. 31).
- Inhorn, Marcia C. and Pasquale Patrizio (2014). “Infertility around the globe: New thinking on gender, reproductive technologies and global movements in the 21st century”. In: *Human Reproduction Update* 21.4, pp. 411–426. DOI: 10.1093/humupd/dmv016 (cit. on p. 19).
- Ji, Jingjuan, Yusheng Liu, Xian Hong Tong, Lihua Luo, Jinlong Ma, and Zijiang Chen (2013). “The optimum number of oocytes in IVF treatment: An analysis of 2455 cycles in China”. In: *Human Reproduction* 28.10, pp. 2728–2734. DOI: 10.1093/humrep/det303 (cit. on pp. 25, 63).
- Jiang, Shancheng, Qize Liu, and Beichen Ding (2023). “A systematic review of the modelling of patient arrivals in emergency departments”. In: *Quantitative Imaging in Medicine and Surgery* 13.3, pp. 1957–1971. DOI: 10.21037/qims-22-268 (cit. on p. 8).
- Jiang, Victoria S., Hemanth Kandula, Prudhvi Thirumalaraju, Manoj Kumar Kanakasabapathy, Panagiotis Cherouveim, Irene Souter, Irene Dimitriadis, Charles L. Bormann, and Hadi Shafiee (2023). “The use of voting ensembles to improve the accuracy of deep neural networks as a non-invasive method to predict embryo ploidy status”. In: *Journal of Assisted Reproduction and Genetics*, pp. 301–308. DOI: 10.1007/s10815-022-02707-6 (cit. on pp. 24, 27).
- Jiang, Victoria S., Deeksha Kartik, Prudhvi Thirumalaraju, Hemanth Kandula, Manoj Kumar Kanakasabapathy, Irene Souter, Irene Dimitriadis, Charles L. Bormann, and Hadi Shafiee (2022). “Advancements in the future of automating micromanipulation techniques in the IVF laboratory using deep convolutional neural networks”. In: *Journal of Assisted Reproduction and Genetics*, pp. 251–257. DOI: 10.1007/s10815-022-02685-9 (cit. on pp. 24, 29).

- Jones, HM, Y Chen, C Gibson, T Heimbach, N Parrott, SA Peters, J Snoeys, VV Upreti, M Zheng, and SD Hall (2015). “Physiologically Based Pharmacokinetic Modelling in Drug Discovery and Development: A Pharmaceutical Industry Perspective”. In: *Clinical Pharmacology and Therapeutics* 97, pp. 247–262 (cit. on p. 13).
- Kakotkin, Viktor V., Ekaterina V. Semina, Tatiana G. Zadorkina, and Mikhail A. Agapov (2023). “Prevention Strategies and Early Diagnosis of Cervical Cancer: Current State and Prospects”. In: *Diagnostics* 13.4, pp. 1–15. DOI: 10.3390/diagnostics13040610 (cit. on p. 7).
- Kallus, Nathan and Angela Zhou (2018). “Policy evaluation and optimization with continuous treatments”. In: *International Conference on Artificial Intelligence and Statistics, AISTATS 2018* 7, pp. 1243–1251 (cit. on p. 16).
- Kamath, Mohan S., Mariano Mascarenhas, Richard Kirubakaran, and Siladitya Bhattacharya (2020). “Number of embryos for transfer following in vitro fertilisation or intra-cytoplasmic sperm injection”. In: *Cochrane Database of Systematic Reviews* 2020.8. DOI: 10.1002/14651858.CD003416.pub5 (cit. on pp. 28, 33, 34, 87, 88, 90, 103).
- Kampaktsis, Polydoros N., Maria Emfietzoglou, Aamna Al Shehhi, Nikolina-Alexia Fasoula, Constantinos Bakogiannis, Dimitrios Mouselimis, Anastasios Tsarouchas, Vassilios P. Vassilikos, Michael Kallmayer, Hans-Henning Eckstein, Leontios Hadjileontiadis, and Angelos Karlas (2023). “Artificial intelligence in atherosclerotic disease: Applications and trends”. In: *Frontiers in Cardiovascular Medicine* 9. January, pp. 1–16. DOI: 10.3389/fcvm.2022.949454 (cit. on p. 7).
- Karl, Kaitlin R., Fermin Jimenez-Krassel, Emily Gibbings, Janet L.H. Ireland, Zaramasina L. Clark, Robert J. Tempelman, Keith E. Latham, and James J. Ireland (2021). “Negative impact of high doses of follicle-stimulating hormone during superovulation on the ovulatory follicle function in small ovarian reserve dairy heifers”. In: *Biology of Reproduction* 104.3, pp. 695–705. DOI: 10.1093/biolre/ioaa210 (cit. on p. 63).
- Kaufman, S J, J L Eastaugh, S Snowden, S W Smye, and V Sharma (1997). “The application of neural networks in predicting the outcome of in-vitro fertilization”. In: *Human reproduction* 12.7, pp. 1454–1457 (cit. on p. 22).
- Keizer, Ron J., Rob ter Heine, Adam Frymoyer, Lawrence J. Lesko, Ranvir Mangat, and Srijib Goswami (2018). “Model-Informed Precision Dosing at the Bedside: Scientific Challenges and Opportunities”. In: *CPT: Pharmacometrics and Systems Pharmacology* 7.12, pp. 785–787. DOI: 10.1002/psp4.12353 (cit. on pp. 12, 13, 57).
- Khan, Bangul, Hajira Fatima, Ayatullah Qureshi, Sanjay Kumar, Abdul Hanan, Jawad Hussain, and Saad Abdullah (2023). “Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector.” In: *Biomedical materials & devices (New York, N.Y.)*, pp. 1–8. DOI: 10.1007/s44174-023-00063-2 (cit. on p. 11).

- Khan-Dawood, F S, L T Goldsmith, G Weiss, and M Y Dawood (1989). “Human corpus luteum secretion of relaxin, oxytocin, and progesterone.” eng. In: *J Clin Endocrinol Metab* 68.3, pp. 627–631. DOI: 10.1210/jcem-68-3-627 (cit. on p. 19).
- Khosravi, Pegah, Ehsan Kazemi, Qiansheng Zhan, Jonas E. Malmsten, Marco Toschi, Pantelis Zisimopoulos, Alexandros Sigaras, Stuart Lavery, Lee A. D. Cooper, Cristina Hickman, Marcos Meseguer, Zev Rosenwaks, Olivier Elemento, Nikica Zaninovic, and Iman Hajirasouliha (2019). “Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization”. In: *npj Digital Medicine* 2.1, pp. 1–9. DOI: 10.1038/s41746-019-0096-y (cit. on pp. 24, 27).
- Koch, Gilbert, Marc Pfister, Imant Daunhawer, Melanie Wilbaux, Sven Wellmann, and Julia E. Vogt (2020). “Pharmacometrics and Machine Learning Partner to Advance Clinical Data Analysis”. In: *Clinical Pharmacology and Therapeutics* 107.4, pp. 926–933. DOI: 10.1002/cpt.1774 (cit. on p. 13).
- Küpker, W., K. Diedrich, and R. G. Edwards (1998). “Principles of mammalian fertilization”. In: *Human Reproduction* 13.SUPPL. 1, pp. 20–32. DOI: 10.1093/humrep/13.suppl\\_.1.20 (cit. on p. 18).
- La Marca, A., E. Papaleo, V. Grisendi, C. Argento, S. Giulini, and A. Volpe (2012). “Development of a nomogram based on markers of ovarian reserve for the individualisation of the follicle-stimulating hormone starting dose in in vitro fertilisation cycles”. In: *BJOG: An International Journal of Obstetrics and Gynaecology* 119.10, pp. 1171–1179. DOI: 10.1111/j.1471-0528.2012.03412.x (cit. on pp. 24, 32, 39, 52, 65).
- Lannon, Benjamin M., Bokyung Choi, Michele R. Hacker, Laura E. Dodge, Beth A. Malizia, C. Brent Barrett, Wing H. Wong, Mylene W.M. Yao, and Alan S. Penzias (2012). “Predicting personalized multiple birth risks after in vitro fertilization-double embryo transfer”. In: *Fertility and Sterility* 98.1, pp. 69–76. DOI: 10.1016/j.fertnstert.2012.04.011 (cit. on pp. 24, 35).
- Lensen, Sarah F., Jack Wilkinson, Jori A. Leijdekkers, Antonio La Marca, Ben Willem J. Mol, Jane Marjoribanks, Helen Torrance, and Frank J. Broekmans (2018). “Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI)”. In: *Cochrane Database of Systematic Reviews* 2018.2. DOI: 10.1002/14651858.CD012693.pub2 (cit. on pp. 31, 63).
- Lenton, E A, B M Landgren, and L Sexton (n.d.). “Normal variation in the length of the luteal phase of the menstrual cycle: identification of the short luteal phase.” eng. In: *Br J Obstet Gynaecol* 91.7 (), pp. 685–689. DOI: 10.1111/j.1471-0528.1984.tb04831.x (cit. on p. 18).
- Lewis, Sheiner and Jon Wakefield (1999). “Modelling in Drug Development”. In: *Statistical Methods in Medical Research* 8.3, pp. 183–193. DOI: <https://doi.org/10.1177/096228029900800302> (cit. on p. 12).

- Liang, Xiaowen, Jiamin Liang, Fengyi Zeng, Yan Lin, Yuewei Li, Kuan Cai, Dong Ni, and Zhiyi Chen (2022). “Evaluation of oocyte maturity using artificial intelligence quantification of follicle volume biomarker by three-dimensional ultrasound”. In: *Reproductive BioMedicine Online* 45.6, pp. 1197–1206. DOI: 10.1016/j.rbmo.2022.07.012 (cit. on pp. 24, 26).
- Lledo, Belen, Jose A. Ortiz, Joaquin Llacer, and Rafael Bernabeu (2014). “Pharmacogenetics of ovarian response”. In: *Pharmacogenomics* 15.6, pp. 885–893. DOI: 10.2217/pgs.14.49 (cit. on pp. 32, 54).
- Luo, Xiu, Li Pei, Yao He, Fujie Li, Wei Han, Shun Xiong, Shubiao Han, Jingyu Li, Xiaodong Zhang, Guoning Huang, and Hong Ye (2022). “High initial FSH dosage reduces the number of available cleavage-stage embryos in a GnRH-antagonist protocol: Real-world data of 8,772 IVF cycles from China”. In: *Frontiers in Endocrinology* 13.October, pp. 1–10. DOI: 10.3389/fendo.2022.986438 (cit. on pp. 25, 63).
- Maggiulli, Roberta, Danilo Cimadomo, Gemma Fabozzi, Letizia Papini, Lisa Dovere, Filippo Maria Ubaldi, and Laura Rienzi (2020). “The effect of ICSI-related procedural timings and operators on the outcome”. In: *Human Reproduction* 35.1, pp. 32–43. DOI: 10.1093/humrep/dez234 (cit. on pp. 32, 63).
- Martínez-Granados, Luis, María Serrano, Antonio González-Utor, Nereyda Ortiz, Vicente Bada-joz, María Luisa López-Regalado, Montserrat Boada, and Jose A. Castilla (2018). “Reliability and agreement on embryo assessment: 5 years of an external quality control programme”. In: *Reproductive BioMedicine Online* 36.3, pp. 259–268. DOI: 10.1016/j.rbmo.2017.12.008 (cit. on p. 26).
- McCallum, Christopher, Jason Riordon, Yihe Wang, Tian Kong, Jae Bem You, Scott Sanner, Alexander Lagunov, Thomas G. Hannam, Keith Jarvi, and David Sinton (2019). “Deep learning-based selection of human sperm with high DNA integrity”. In: *Communications Biology* 2.1, pp. 1–10. DOI: 10.1038/s42003-019-0491-6 (cit. on pp. 24, 29).
- McComb, Mason, Robert Bies, and Murali Ramanathan (2022). “Machine learning in pharmacometrics: Opportunities and challenges”. In: *British Journal of Clinical Pharmacology* 88.4, pp. 1482–1499. DOI: 10.1111/bcp.14801 (cit. on p. 13).
- McComb, Mason and Murali Ramanathan (2020). “Generalized Pharmacometric Modeling, a Novel Paradigm for Integrating Machine Learning Algorithms: A Case Study of Metabolomic Biomarkers”. In: *Clinical Pharmacology and Therapeutics* 107.6, pp. 1343–1351. DOI: 10.1002/cpt.1746 (cit. on p. 13).
- Medenica, Sanja, Dusan Zivanovic, Ljubica Batkoska, Susanna Marinelli, Giuseppe Basile, Antonio Perino, Gaspare Cucinella, Giuseppe Gullo, and Simona Zaami (2022). “The Future Is Coming: Artificial Intelligence in the Treatment of Infertility Could Improve Assisted Reproduction Outcomes—The Value of Regulatory Frameworks”. In: *Diagnostics* 12.12. DOI: 10.3390/diagnostics12122979 (cit. on p. 23).



- Melarkode, Navneet, Kathiravan Srinivasan, Saeed Mian Qaisar, and Pawel Plawiak (2023). “AI-Powered Diagnosis of Skin Cancer: A Contemporary Review, Open Challenges and Future Research Directions.” eng. In: *Cancers (Basel)* 15.4. DOI: 10.3390/cancers15041183 (cit. on p. 7).
- Mendizabal-Ruiz, Gerardo, Alejandro Chavez-Badiola, Isaac Aguilar Figueroa, Vladimir Martinez Nuño, Adolfo Flores-Saiffe Farias, Roberto Valencia-Murilloa, Andrew Drakeley, Juan Paulo Garcia-Sandoval, and Jacques Cohen (2022). “Computer software (SiD) assisted real-time single sperm selection associated with fertilization and blastocyst formation”. In: *Reproductive BioMedicine Online* 45.4, pp. 703–711. DOI: 10.1016/j.rbmo.2022.03.036 (cit. on pp. 24, 29).
- Milewski, Robert, Agnieszka Kuczyńska, Bożena Stankiewicz, and Waldemar Kuczyński (2017). “How much information about embryo implantation potential is included in morphokinetic data? A prediction model based on artificial neural networks and principal component analysis”. In: *Advances in Medical Sciences* 62.1, pp. 202–206. DOI: 10.1016/j.advms.2017.02.001 (cit. on pp. 24, 27).
- Mirroshandel, Seyed Abolghasem, Fatemeh Ghasemian, and Sara Monji-Azad (2016). “Applying data mining techniques for increasing implantation rate by selecting best sperms for intracytoplasmic sperm injection treatment”. In: *Computer Methods and Programs in Biomedicine* 137, pp. 215–229. DOI: 10.1016/j.cmpb.2016.09.013 (cit. on pp. 24, 28).
- Mirsky, Simcha K., Itay Barnea, Mattan Levi, Hayit Greenspan, and Natan T. Shaked (2017). “Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning”. In: *Cytometry Part A* 91.9, pp. 893–900. DOI: 10.1002/cyto.a.23189 (cit. on pp. 24, 28).
- Miyagi, Yasunari, Toshihiro Habara, Rei Hirata, and Nobuyoshi Hayashi (2019). “Feasibility of deep learning for predicting live birth from a blastocyst image in patients classified by age”. In: *Reproductive Medicine and Biology* 18.2, pp. 190–203. DOI: 10.1002/rmb2.12266 (cit. on pp. 24, 27).
- Mohr-Sasson, Aya, Raoul Orvieto, Shlomit Blumenfeld, Michal Axelrod, Danielle Mor-Hadar, Leonti Grin, Adva Aizer, and Jigal Haas (2020). “The association between follicle size and oocyte development as a function of final follicular maturation triggering”. In: *Reproductive BioMedicine Online* 40.6, pp. 887–893. DOI: 10.1016/j.rbmo.2020.02.005 (cit. on p. 25).
- Morales, Dinora A, Endika Bengoetxea, and Pedro Larrañaga (2008). “Selection of human embryos for transfer by Bayesian classifiers.” eng. In: *Comput Biol Med* 38.11-12, pp. 1177–1186. DOI: 10.1016/j.combiomed.2008.09.002 (cit. on p. 27).
- Motwani, Anand, Piyush Kumar, and Mahesh Pawar (2020). “Ubiquitous and smart healthcare monitoring frameworks based on machine learning: A comprehensive review”. In: January (cit. on p. 8).

- Naether, Olaf G.J., Andreas Tandler-Schneider, and Wilma Bilger (2015). “Individualized recombinant human follicle-stimulating hormone dosing using the CONSORT calculator in assisted reproductive technology: A large, multicenter, observational study of routine clinical practice”. In: *Drug, Healthcare and Patient Safety* 7, pp. 69–76. DOI: 10.2147/DHPS.S77320 (cit. on p. 33).
- Nazir, Talha, Muhammad Mushhood Ur Rehman, Muhammad Roshan Asghar, and Junaid S. Kalia (2022). “Artificial intelligence assisted acute patient journey”. In: *Frontiers in Artificial Intelligence* 5. DOI: 10.3389/frai.2022.962165 (cit. on p. 8).
- Nelson, Scott M., Richard Fleming, Marco Gaudoin, Bokyoung Choi, Kenny Santo-Domingo, and Mylene Yao (2015). “Antimüllerian hormone levels and antral follicle count as prognostic indicators in a personalized prediction model of live birth”. In: *Fertility and Sterility* 104.2, pp. 325–332. DOI: 10.1016/j.fertnstert.2015.04.032 (cit. on pp. 24, 29).
- Neves, A. R., N. L. Vuong, C. Blockeel, S. Garcia, C. Alviggi, C. Spits, P. Q.M. Ma, M. T. Ho, H. Tournaye, and N. P. Polyzos (2022). “The effect of polymorphisms in FSHR gene on late follicular phase progesterone and estradiol serum levels in predicted normoresponders”. In: *Human Reproduction* 37.11, pp. 2646–2654. DOI: 10.1093/humrep/deac193 (cit. on p. 32).
- Nyboe Andersen, Anders et al. (2017). “Individualized versus conventional ovarian stimulation for in vitro fertilization: a multicenter, randomized, controlled, assessor-blinded, phase 3 non-inferiority trial”. In: *Fertility and Sterility* 107.2, 387–396.e4. DOI: 10.1016/j.fertnstert.2016.10.033 (cit. on pp. 24, 32, 37, 39, 52, 54, 74).
- Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan (2019). “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366.6464, pp. 447–453. DOI: 10.1126/science.aax2342 (cit. on pp. 10, 11, 94).
- Oh, Sam S, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E Barcelo, Marquitta J White, Danielle M de Bruin, Ruth M Greenblatt, Kirsten Bibbins-Domingo, Alan H B Wu, Luisa N Borrell, Chris Gunter, Neil R Powe, and Esteban G Burchard (2015). “Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled.” eng. In: *PLoS Med* 12.12, e1001918. DOI: 10.1371/journal.pmed.1001918 (cit. on p. 57).
- Olivennes, F., G. Trew, A. Borini, F. Broekmans, P. Arriagada, D. W. Warne, and C. M. Howles (2015). “Randomized, controlled, open-label, non-inferiority study of the CONSORT algorithm for individualized dosing of follitropin alfa”. In: *Reproductive BioMedicine Online* 30.3, pp. 248–257. DOI: 10.1016/j.rbmo.2014.11.013 (cit. on pp. 24, 33, 37, 39, 52, 54, 74).
- Ory, Jesse, Michael B. Tradewell, Udi Blankstein, Thiago F. Lima, Sirpi Nackeeran, Daniel C. Gonzalez, Elie Nwefo, Joseph Moryousef, Vinayak Madhusoodanan, Susan Lau, Keith Jarvi, and Ranjith Ramasamy (2022). “Artificial Intelligence Based Machine Learning Models Predict Sperm parameter Upgrading after Varicocele Repair: A Multi-Institutional Analysis”. In: *World Journal of Men’s Health* 40.4, pp. 618–626. DOI: 10.5534/wjmh.210159 (cit. on pp. 24, 30).

- OS Guideline Development Group (2019). “Ovarian Stimulation for IVF/ICSI”. In: *European Society of Human Reproduction and Embryology (ESHRE)* October, pp. 1–136 (cit. on p. 40).
- Ota, Ryosaku and Fumiyoshi Yamashita (2022). “Application of machine learning techniques to the analysis and prediction of drug pharmacokinetics”. In: *Journal of Controlled Release* 352.November, pp. 961–969. DOI: 10.1016/j.jconrel.2022.11.014 (cit. on p. 17).
- Pache, T. D., J. W. Wladimiroff, F. H. De Jong, W. C. Hop, and B. C.J.M. Fauser (1990). “Growth patterns of nondominant ovarian follicles during the normal menstrual cycle”. In: *Fertility and Sterility* 54.4, pp. 638–642. DOI: 10.1016/S0015-0282(16)53821-7 (cit. on p. 18).
- Paternot, Goedeke, Alex M. Wetsels, Fabienne Thonon, Anne Vansteenbrugge, Dorien Willemen, Johanna Devroe, Sophie Debrock, Thomas M. D’Hooghe, and Carl Spiessens (2011). “Intra- and interobserver analysis in the morphological assessment of early stage embryos during an IVF procedure: A multicentre study”. In: *Reproductive Biology and Endocrinology* 9.1, p. 127. DOI: 10.1186/1477-7827-9-127 (cit. on p. 26).
- Pearl, Judea (2000). “Why there is no statistical test for confounding, why many think there is, and why they are almost right”. In: *Causality: Models, Reasoning, and Inference*. January. Cambridge University Press. Chap. 6 (cit. on pp. 14, 15).
- (2010). “An Introduction to Causal Inference”. In: *The international journal of biostatistics* 6.2, Article 7 (cit. on p. 12).
- Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons, p. 160 (cit. on pp. 14–16).
- Pépin, Jean Louis, Bruno Degano, Renaud Tamisier, and Damien Viglino (2022). “Remote Monitoring for Prediction and Management of Acute Exacerbations in Chronic Obstructive Pulmonary Disease (AECOPD)”. In: *Life* 12.4. DOI: 10.3390/life12040499 (cit. on p. 8).
- Permadi, Wiryawan, Mohammad Wahyu Ferdian, Dian Tjahyadi, Wulan Ardhana Iswari, and Tono Djuwantono (2021). “Correlation of anti-mullerian hormone level and antral follicle count with oocyte number in a fixed-dose controlled ovarian hyperstimulation of patients of In Vitro fertilization program”. In: *International Journal of Fertility and Sterility* 15.1, pp. 40–43. DOI: 10.22074/IJFS.2021.6238 (cit. on p. 25).
- Pesapane, Filippo et al. (2023). “How Radiomics Can Improve Breast Cancer Diagnosis and Treatment”. In: *Journal of Clinical Medicine* 12.4. DOI: 10.3390/jcm12041372 (cit. on p. 7).
- Petersen, Bjørn Molt, Mikkel Boel, Markus Montag, and David K. Gardner (2016). “Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3”. In: *Human Reproduction* 31.10, pp. 2231–2244. DOI: 10.1093/humrep/dew188 (cit. on pp. 24, 27).
- Polyzos, N. P. and S. K. Sunkara (2015). “Sub-optimal responders following controlled ovarian stimulation: An overlooked group?” In: *Human Reproduction* 30.9, pp. 2005–2008. DOI: 10.1093/humrep/dev149 (cit. on pp. 40, 63, 64).

- Polyzos, Nikolaos P., A. R. Neves, P. Drakopoulos, C. Spits, B. Alvaro Mercadal, S. Garcia, P. Q.M. Ma, L. H. Le, M. T. Ho, J. Mertens, D. Stoop, H. Tournaye, and N. L. Vuong (2021). “The effect of polymorphisms in FSHR and FSHB genes on ovarian response: a prospective multicenter multinational study in Europe and Asia”. In: *Human Reproduction* 36.6, pp. 1711–1721. DOI: 10.1093/humrep/deab068 (cit. on p. 32).
- Porchet, H. C., J. Y. Le Cotonnec, and E. Loumaye (1994). “Clinical pharmacology of recombinant human follicle-stimulating hormone. III. Pharmacokinetic-pharmacodynamic modeling after repeated subcutaneous administration”. In: *Fertility and Sterility* 61.4, pp. 687–695. DOI: 10.1016/s0015-0282(16)56646-1 (cit. on pp. 30, 63).
- Potlitz, F., A. Link, and L. Schulig (2023). “Advances in the discovery of new chemotypes through ultra-large library docking”. In: *Expert Opinion on Drug Discovery* 18.3, pp. 303–314. DOI: 10.1080/17460441.2023.2171984 (cit. on p. 8).
- Pouly, Jean Luc, François Olivennes, Nathalie Massin, Médéric Celle, Natacha Caizergues, and Francis Contard (2015). “Usability and utility of the CONSORT calculator for FSH starting doses: A prospective observational study”. In: *Reproductive BioMedicine Online*. DOI: 10.1016/j.rbmo.2015.06.001 (cit. on p. 33).
- Poweleit, Ethan A, Alexander A Vinks, and Tomoyuki Mizuno (2023). “Artificial Intelligence and Machine Learning Approaches to Facilitate Therapeutic Drug Management and Model-Informed Precision Dosing.” eng. In: *Ther Drug Monit* 45.2, pp. 143–150. DOI: 10.1097/FTD.0000000000001078 (cit. on p. 17).
- Powles, Julia and Hal Hodson (2017). “Google DeepMind and healthcare in an age of algorithms”. en. In: *Health and Technology* 7.4, pp. 351–367. DOI: 10.1007/s12553-017-0179-1 (cit. on p. 10).
- Pujol, A, O Cairó, T Mukan, V Pérez, D García, R Vassena, and D Mataró (2021). “P-668 We aim for one baby, not one embryo: a personalized ET strategy based on embryo score and woman age maximizes LB and minimizes twins”. In: *Human Reproduction* 36.Supplement.1. deab130.667. DOI: 10.1093/humrep/deab130.667 (cit. on p. 104).
- Raine-Fenning, Nicholas J., Bruce K. Campbell, Jeanette S. Clewes, Nigel R. Kendall, and Ian R. Johnson (2004). “Defining endometrial growth during the menstrual cycle with three-dimensional ultrasound”. In: *BJOG: An International Journal of Obstetrics and Gynaecology* 111.9, pp. 944–949. DOI: 10.1111/j.1471-0528.2004.00214.x (cit. on p. 18).
- Ranjbari, Sima, Toktam Khatibi, Ahmad Vosough Dizaji, Hesamoddin Sajadi, Mehdi Totonchi, and Firouzeh Ghaffari (2021). “CNFE-SE: a novel approach combining complex network-based feature engineering and stacked ensemble to predict the success of intrauterine insemination and ranking the features”. In: *BMC Medical Informatics and Decision Making* 21.1, pp. 1–29. DOI: 10.1186/s12911-020-01362-0 (cit. on pp. 24, 29).
- Raza, Muhammad Ahmer, Shireen Aziz, Misbah Noreen, Amna Saeed, Irfan Anjum, Mudassar Ahmed, and Shahid Masood Raza (2022). “Artificial Intelligence (AI) in Pharmacy: An

- Overview of Innovations”. In: *INNOVATIONS in pharmacy* 13.2, p. 13. DOI: 10.24926/iip.v13i2.4839 (cit. on p. 8).
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “”Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Aug, pp. 1135–1144. DOI: 10.1145/2939672.2939778 (cit. on p. 10).
- Richards, J. S. (1980). “Maturation of ovarian follicles: actions and interactions of pituitary and ovarian hormones on follicular cell differentiation”. In: *Physiological Reviews* 60.1, pp. 51–89. DOI: 10.1152/physrev.1980.60.1.51 (cit. on p. 30).
- Roberts, S. A., W. M. Hirst, D. R. Brison, and A. Vail (2010). “Embryo and uterine influences on IVF outcomes: An analysis of a UK multi-centre cohort”. In: *Human Reproduction* 25.11, pp. 2792–2802. DOI: 10.1093/humrep/deq213 (cit. on pp. 24, 34).
- Roberts, S. A., L. McGowan, W. M. Hirst, D. R. Brison, A. Vail, and B. A. Lieberman (2010). “Towards single embryo transfer? modelling clinical outcomes of potential treatment choices using multiple data sources: Predictive models and patient perspectives”. In: *Health Technology Assessment* 14.38, pp. 1–237. DOI: 10.3310/hta14380 (cit. on pp. 24, 34, 35, 88, 94).
- Roberts, Stephen A. (2007). “Models for assisted conception data with embryo-specific covariates”. In: *Statistics in Medicine* 26.1, pp. 156–170. DOI: 10.1002/sim.2525 (cit. on pp. 24, 27, 34).
- Roberts, Stephen A., Linda McGowan, W. Mark Hirst, Andy Vail, Anthony Rutherford, Brian A. Lieberman, and Daniel R. Brison (2011). “Reducing the incidence of twins from IVF treatments: Predictive modelling from a retrospective cohort”. In: *Human Reproduction* 26.3, pp. 569–575. DOI: 10.1093/humrep/deq352 (cit. on pp. 24, 35, 88).
- Roberts, Stephen A. and Christos Stylianou (2012). “The non-independence of treatment outcomes from repeat IVF cycles: Estimates and consequences”. In: *Human Reproduction* 27.2, pp. 436–443. DOI: 10.1093/humrep/der420 (cit. on pp. 24, 27).
- Rozario, Natasha and Duncan Rozario (2020). “Can machine learning optimize the efficiency of the operating room in the era of COVID-19?” In: *Canadian Journal of Surgery* 63.6, E537–E529. DOI: 10.1503/CJS.016520 (cit. on p. 8).
- Ruiz-Alonso, Maria, David Blesa, Patricia Díaz-Gimeno, Eva Gómez, Manuel Fernández-Sánchez, Francisco Carranza, Joan Carrera, Felip Vilella, Antonio Pellicer, and Carlos Simón (2013). “The endometrial receptivity array for diagnosis and personalized embryo transfer as a treatment for patients with repeated implantation failure”. In: *Fertility and Sterility* 100.3, pp. 818–824. DOI: 10.1016/j.fertnstert.2013.05.004 (cit. on p. 28).
- Sacks, Bill, Gregory Meyerson, and Jeffry A. Siegel (2016). “Epidemiology Without Biology: False Paradigms, Unfounded Assumptions, and Specious Statistics in Radiation Science (with Commentaries by Inge Schmitz-Feuerhake and Christopher Busby and a Reply by the Au-

- thors)". In: *Biological Theory* 11.2, pp. 69–101. DOI: 10.1007/s13752-016-0244-4 (cit. on p. 11).
- Sarkar, Chayna, Biswadeep Das, Vikram Singh Rawat, Julie Birdie Wahlang, Arvind Nongpiur, Iadarilang Tiewsoh, Nari M Lyngdoh, Debasmita Das, Manjunath Bidarolli, and Hannah Theresa Sony (2023). "Artificial Intelligence and Machine Learning Technology Driven Modern Drug Discovery and Development". In: *International Journal of Molecular Sciences* 24.3, p. 2026. DOI: 10.3390/ijms24032026 (cit. on p. 8).
- Scholkopf, Bernhard, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio (2021). "Toward Causal Representation Learning". In: *Proceedings of the IEEE* 109.5, pp. 612–634. DOI: 10.1109/JPROC.2021.3058954 (cit. on pp. 14, 16).
- Schwab, Patrick, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen (2019). "Learning Counterfactual Representations for Estimating Individual Dose-Response Curves". In: (cit. on p. 16).
- Shahrokh Tehraninezhad, Ensieh, Fatemeh Mehrabi, Raheleh Taati, Vahid Kalantar, Elham Azimineko, and Azam Tarafdari (2016). *Analysis of ovarian reserve markers (AMH, FSH, AFC) in different age strata in IVF/ICSI patients*. Tech. rep. 8, pp. 501–506 (cit. on p. 31).
- Shao, Jun, Jiechao Ma, Qin Zhang, Weimin Li, and Chengdi Wang (2023). "Seminars in Cancer Biology Predicting gene mutation status via artificial intelligence technologies based on multimodal integration (MMI) to advance precision oncology". In: *Seminars in Cancer Biology* 91.August 2022, pp. 1–15. DOI: 10.1016/j.semcancer.2023.02.006 (cit. on p. 8).
- Sheiner, L B and J-L Steimer (2000). "Pharmacokinetic/Pharmacodynamic Modeling in Drug Development". In: *Annual Review of Pharmacology and Toxicology* 40.1, pp. 67–95. DOI: 10.1146/annurev.pharmtox.40.1.67 (cit. on p. 13).
- Sheiner, Lewis B. and Stuart L. Beal (1982). "Bayesian individualization of pharmacokinetics: Simple implementation and comparison with non-Bayesian methods". In: *Journal of Pharmaceutical Sciences* 71.12, pp. 1344–1348. DOI: 10.1002/jps.2600711209 (cit. on p. 13).
- Sheiner, Lewis B. and Thomas M. Ludden (1992). "Population pharmacokinetics/dynamics". In: *Annual Review of Pharmacology and Toxicology* 32, pp. 185–209. DOI: 10.1146/annurev.pa.32.040192.001153 (cit. on p. 13).
- Simopoulou, Mara, Konstantinos Sfakianoudis, Evangelos Maziotis, Nikolaos Antoniou, Anna Rapani, George Anifandis, Panagiotis Bakas, Stamatis Bolaris, Agni Pantou, Konstantinos Pantos, and Michael Koutsilieris (2018). "Are computational applications the "crystal ball" in the IVF laboratory? The evolution from mathematics to artificial intelligence". In: *Journal of Assisted Reproduction and Genetics* 35.9, pp. 1545–1557. DOI: 10.1007/s10815-018-1266-6 (cit. on p. 23).
- Simpson, E. H. (1951). "The Interpretation of Interaction in Contingency Tables". In: *Journal of the Royal Statistical Society. Series B, Methodological* 13.2, pp. 238–241 (cit. on p. 14).

- Speirs, A. L., A. Lopata, M. J. Gronow, G. N. Kellow, and W. I. Johnston (1983). “Analysis of the benefits and risks of multiple embryo transfer”. In: *Fertility and Sterility* 39.4, pp. 468–471. DOI: 10.1016/S0015-0282(16)46933-5 (cit. on p. 27).
- Stan-Ilie, Madalina, Vasile Sandru, Gabriel Constantinescu, Oana Mihaela Plotogea, Ecaterina Mihaela Rinja, Iulia Florentina Tincu, Alexandra Jichitu, Adriana Elena Carasel, Andreea Cristina Butuc, and Bogdan Popa (2023). “Artificial Intelligence—The Rising Star in the Field of Gastroenterology and Hepatology”. In: *Diagnostics* 13.4, pp. 1–15. DOI: 10.3390/diagnostics13040662 (cit. on p. 7).
- Steiner, Anne Z., David Pritchard, Frank Z. Stanczyk, James S. Kesner, Juliana W. Meadows, Amy H. Herring, and Donna D. Baird (2017). “Association between biomarkers of ovarian reserve and infertility among older women of reproductive age”. In: *JAMA - Journal of the American Medical Association* 318.14, pp. 1367–1376. DOI: 10.1001/jama.2017.14588 (cit. on p. 31).
- Steward, Ryan G., Lan Lan, Anish A. Shah, Jason S. Yeh, Thomas M. Price, James M. Goldfarb, and Suheil J. Muasher (2014). “Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: An analysis of 256,381 in vitro fertilization cycles”. In: *Fertility and Sterility* 101.4, pp. 967–973. DOI: 10.1016/j.fertnstert.2013.12.026 (cit. on pp. 25, 40, 63).
- Storr, Ashleigh, Christos A. Venetis, Simon Cooke, Suha Kilani, and William Ledger (2017). “Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: A multicenter study”. In: *Human Reproduction* 32.2, pp. 307–314. DOI: 10.1093/humrep/dew330 (cit. on p. 26).
- Sun, Bo, Yujia Ma, Lu Li, Linli Hu, Fang Wang, Yile Zhang, Shanjun Dai, and Yingpu Sun (2021). “Factors Associated with Ovarian Hyperstimulation Syndrome (OHSS) Severity in Women With Polycystic Ovary Syndrome Undergoing IVF/ICSI”. In: *Frontiers in Endocrinology* 11.January, pp. 1–8. DOI: 10.3389/fendo.2020.615957 (cit. on p. 31).
- Sunkara, Sesh Kamal, Vivian Rittenberg, Nick Raine-Fenning, Siladitya Bhattacharya, Javier Zamora, and Arri Coomarasamy (2011). “Association between the number of eggs and live birth in IVF treatment: An analysis of 400 135 treatment cycles”. In: *Human Reproduction* 26.7, pp. 1768–1774. DOI: 10.1093/humrep/der106 (cit. on pp. 25, 40, 63).
- The ESHRE Guideline Group on Ovarian Stimulation et al. (2020). “ESHRE guideline: ovarian stimulation for IVF/ICSI†”. In: *Human Reproduction Open* 2020.2. DOI: 10.1093/hropen/hoaa009 (cit. on pp. 32, 68).
- Theodorou, Efsthios, Benjamin P. Jones, Suzanne Cawood, Carleen Heath, Paul Serhal, and Jara Ben-Nagi (2021). “Adding a low-quality blastocyst to a high-quality blastocyst for a double embryo transfer does not decrease pregnancy and live birth rate”. In: *Acta Obstetrica et Gynecologica Scandinavica* 100.6, pp. 1124–1131. DOI: 10.1111/aogs.14088 (cit. on p. 104).
- Tran, D., S. Cooke, P. J. Illingworth, and D. K. Gardner (2019). “Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer”. In: *Human Reproduction* 34.6, pp. 1011–1018. DOI: 10.1093/humrep/dez064 (cit. on pp. 24, 27).

- Vaegter, Katarina Kebbon, Lars Berglund, Johanna Tilly, Nermin Hadziosmanovic, Thomas Brodin, and Jan Holte (2019). “Construction and validation of a prediction model to minimize twin rates at preserved high live birth rates after IVF”. In: *Reproductive BioMedicine Online* 38.1, pp. 22–29. DOI: 10.1016/j.rbmo.2018.09.020 (cit. on pp. 24, 35, 88).
- Vaiarelli, Alberto, Danilo Cimadomo, Alessandro Conforti, Mauro Schimberni, Maddalena Giuliani, Pietro D’Alessandro, Silvia Colamaria, Carlo Alviggi, Laura Rienzi, and Filippo Maria Ubaldi (2020). “Luteal phase after conventional stimulation in the same ovarian cycle might improve the management of poor responder patients fulfilling the Bologna criteria: a case series”. In: *Fertility and Sterility* 113.1, pp. 121–130. DOI: 10.1016/j.fertnstert.2019.09.012 (cit. on pp. 32, 63).
- Van Santbrink, E. J.P., W. C. Hop, T. J.H.M. Van Dessel, F. H. De Jong, and B. C.J.M. Fauser (1995). “Decremental follicle-stimulating hormone and dominant follicle development during the normal menstrual cycle”. In: *Fertility and Sterility* 64.1, pp. 37–43. DOI: 10.1016/s0015-0282(16)57652-3 (cit. on p. 18).
- VerMilyea, M., J. M.M. Hall, S. M. Diakiw, A. Johnston, T. Nguyen, D. Perugini, A. Miller, A. Picou, A. P. Murphy, and M. Perugini (2020). “Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF”. In: *Human reproduction (Oxford, England)* 35.4, pp. 770–784. DOI: 10.1093/humrep/deaa013 (cit. on pp. 24, 27).
- Visco, Valeria et al. (2022). “Artificial Intelligence in Hypertension management : an ace up your sleeve .” In: Cv, pp. 1–14 (cit. on p. 8).
- Wen, Jen Yu, Chung Fen Liu, Ming Ting Chung, and Yung Chieh Tsai (2022). “Artificial intelligence model to predict pregnancy and multiple pregnancy risk following in vitro fertilization-embryo transfer (IVF-ET)”. In: *Taiwanese Journal of Obstetrics and Gynecology* 61.5, pp. 837–846. DOI: 10.1016/j.tjog.2021.11.038 (cit. on pp. 24, 36, 88).
- Wilcox, Allen J., Donna Day Baird, and Clarice R. Weinberg (1999). “Time of Implantation of the Conceptus and Loss of Pregnancy”. In: *Obstetrical & Gynecological Survey* 54.11, p. 705. DOI: 10.1097/00006254-199911000-00018 (cit. on p. 28).
- Wilcox, Allen J., Clarice R. Weinberg, and Donna D. Baird (1996). “Timing of Sexual Intercourse in Relation to Ovulation”. In: *Obstetrical & Gynecological Survey* 51.6, pp. 357–358. DOI: 10.1097/00006254-199606000-00016 (cit. on p. 19).
- Wilkosz, Pawel, Gareth D. Greggains, Tom G. Tanbo, and Peter Fedorcsak (2014). “Female reproductive decline is determined by remaining ovarian reserve and age”. In: *PLoS ONE* 9.10. DOI: 10.1371/journal.pone.0108343 (cit. on p. 31).
- Wintner, Eliana Muskin, Anat Hershko-Klement, Keren Tzadikévitch, Yehudith Ghetler, Ofer Gonen, Oren Wintner, Adrian Shulman, and Amir Wiser (2017). “Does the transfer of a poor quality embryo together with a good quality embryo affect the In Vitro Fertilization (IVF)



- outcome?” In: *Journal of Ovarian Research* 10.1, pp. 1–5. DOI: 10.1186/s13048-016-0297-9 (cit. on p. 104).
- World Health Organization (WHO) (2018). *International Classification of Diseases*. <https://icd.who.int/en>. WHO: Geneva, Switzerland (cit. on p. 19).
- Wright, Stephen J (2015). “Coordinate descent algorithms”. In: *Mathematical Programming* 151.1, pp. 3–34 (cit. on p. 62).
- Xu, Qian, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijing Yang, Huan Xiong, Yi Lyu, Yue Zhou, and Aijing Luo (2023). “Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review”. In: *Journal of healthcare engineering* 2023, p. 9919269. DOI: 10.1155/2023/9919269 (cit. on p. 10).
- Yang, Chan Yun, Chamani Shiranthika, Chung Yih Wang, Kuo Wei Chen, and Sagara Sumathipala (2023). “Reinforcement learning strategies in cancer chemotherapy treatments: A review”. In: *Computer Methods and Programs in Biomedicine* 229, p. 107280. DOI: 10.1016/j.cmpb.2022.107280 (cit. on p. 17).
- Yoo, Jiho, Tae Yong Kim, Insuk Joung, and Sang Ok Song (2023). “ScienceDirect Structural Biology Industrializing AI / ML during the end-to-end drug discovery process”. In: *Current Opinion in Structural Biology* 79, p. 102528. DOI: 10.1016/j.sbi.2023.102528 (cit. on p. 8).
- Zeadna, A., N. Khateeb, L. Rokach, Y. Lior, I. Har-Vardi, A. Harlev, M. Huleihel, E. Lunenfeld, and E. Levitas (2020). “Prediction of sperm extraction in non-obstructive azoospermia patients: A machine-learning perspective”. In: *Human Reproduction* 35.7, pp. 1505–1514. DOI: 10.1093/humrep/deaa109 (cit. on pp. 24, 30).
- Zhang, Wen bi, Qing Li, Hu Liu, Wei jian Chen, Chun lei Zhang, He Li, Xiang Lu, Jun ling Chen, Lu Li, Han Wu, and Xiao xi Sun (2021). “Transcriptomic analysis of endometrial receptivity for a genomic diagnostics model of Chinese women”. In: *Fertility and Sterility* 116.1, pp. 157–164. DOI: 10.1016/j.fertnstert.2020.11.010 (cit. on pp. 24, 28).
- Zhou, H and C R Weinberg (1998). “Evaluating effects of exposures on embryo viability and uterine receptivity in in vitro fertilization.” eng. In: *Stat Med* 17.14, pp. 1601–1612. DOI: 10.1002/(sici)1097-0258(19980730)17:14<1601::aid-sim870>3.0.co;2-2 (cit. on p. 27).
- Zhu, Qianqian, Jiaying Lin, Haoyuan Gao, Ningling Wang, Bian Wang, and Yun Wang (2020). “The Association Between Embryo Quality, Number of Transferred Embryos and Live Birth Rate After Vitrified Cleavage-Stage Embryos and Blastocyst Transfer”. In: *Frontiers in Physiology* 11.August, pp. 1–7. DOI: 10.3389/fphys.2020.00930 (cit. on p. 104).