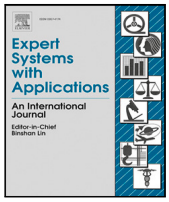
Expert Systems With Applications 238 (2024) 121796

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



IDoser: Improving individualized dosing policies with clinical practice and machine learning

Nuria Correa [a,b,c], Jesus Cerquides [b,*], Rita Vassena [c,1], Mina Popovic [c], Josep Lluis Arcos [b]

[a] Universitat Autonoma de Barcelona (UAB), Bellaterra, Barcelona, 08193, Spain
[b] Artificial Intelligence Research Institute, IIIA (CSIC), Campus de la UAB, Bellaterra, Barcelona, 08193, Spain
[c] Clinica Eugin, Carrer de Balmes 236, Barcelona, 08006, Spain

ARTICLE INFO

Keywords:
Decision support system
Ovarian stimulation
Individualized dosing
Observational datasets
FSH

ABSTRACT

Optimizing drug dosages is essential for effective treatment. Clinical protocols may not suit all types of patients evenly, due to many drug trials not being designed to account for all comorbities or clinically relevant outcomes. Methodologies to optimize drug policies with observational data exist, but struggle due to limited data completeness in clinical settings. Computational methods can help overcome these challenges by leveraging field knowledge.

This paper proposes an Individualized Doser (IDoser), a core dosing model that links drug dose to relevant covariates via a set of coefficients and includes a loss function to code needed assumptions and requirements. Coordinate descent is used to obtain a fitted model with minimal loss. The loss function also measures performance when validating the model with unseen data. We validated the proposed approach using the case of follicle-stimulating hormone (FSH) dosing for controlled ovarian stimulation (COS).

When compared to clinical practice, IDoser achieved a net improvement of up to 31.97% in the validation cases.

We present a simple but effective method to bridge the gap between current clinical dosing policies and gold policies based on the true underlying and often unknown dose–response functions.

1. Introduction

Once a patient entrusts their health to a clinical professional for treatment, they have every right to expect the highest standard of care. Among the numerous essential steps involved in a particular treatment, one critical aspect involves the accurate determination of the appropriate dose for a prescribed medication. This dose will be determined with an optimal outcome in mind, such as maintaining the patient's blood pressure within a balanced range or ensuring that their temperature falls within specific values. Clinical professionals rely on available knowledge of the underlying dose–response relationship to determine the appropriate drug dose to achieve optimal outcomes. However, this relationship is often not well-understood, which may result in suboptimal dose selection, ultimately compromising clinical care for the patient.

An example of this situation can be found in the empirical case discussed in this research, which pertains to the challenge of determining the optimal first dose of follicle-stimulating hormone (FSH) for controlled ovarian stimulation (COS). COS is a key step in the treatment of infertility. It is used to induce the ovary to develop multiple follicles and eggs (oocytes) simultaneously. Once the follicles reach the appropriate size, they are punctured and aspirated in a simple surgical procedure, leading to the collection of mature oocytes. An appropriate COS is critical to the success of in vitro fertilization (IVF), as the number of mature oocytes retrieved is tightly associated with the chances of achieving pregnancy safely. The desired outcome involves obtaining a specific range of mature oocytes, typically between 10 to 15, as supported by research (Polyzos & Sunkara, 2015; Steward et al., 2014). Deviating from this range, whether too low or too high, is considered undesirable. Specifically, retrieving a lower number of mature oocytes reduces the chances of achieving a successful pregnancy, while a higher number of mature oocytes increases the risk of ovarian hyperstimulation syndrome (OHSS), a serious adverse complication resulting from an exaggerated response to excess hormones.

Individuals with similar characteristics (sometimes even the same individual at different points in time) can respond differently to the

* Corresponding author.
E-mail addresses: ncorrea@eugin.es (N. Correa), cerquide@iiia.csic.es (J. Cerquides), vassena@fecundis.com (R. Vassena), mpopovic@eugin.es (M. Popovic), arcos@iiia.csic.es (J.L. Arcos).
[1] Current adress: Fecundis, Baldiri i Reixac, Barcelona, Spain.

https://doi.org/10.1016/j.eswa.2023.121796
Received 3 May 2023; Received in revised form 12 September 2023; Accepted 21 September 2023
Available online 4 October 2023
0957-4174/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

same treatment – a direct consequence of the lack of knowledge surrounding the FSH dose–response relationship, as well as its interaction with unknown factors.

Pharmacokinetic/pharmacodynamic (PK/PD) studies constitute an important step in drug development. These studies inform the design and execution of Randomized Controlled Trials (RCTs) in which the drug is tested in clinical conditions. Clinical protocols are then derived from these results to guide practice. Ultimately, results obtained from this approach can vary from acceptable to almost optimal, depending on the extent to which the PK/PD models align with the characteristics of the target population and real-world clinical outcomes. While this approach enables most patients to receive an appropriate drug dose, dosing protocols or policies may not cater to all patients equally. This primarily stems from the distribution of the target population and the presence of accompanying comorbidities.

Clinicians often rely on their personal experience and available published literature to compensate for these knowledge gaps. However, this approach may still be suboptimal. PK/PD models can be adapted for real-world clinical use, but either imply the use of new prospective data or very diverse observational datasets, which we define as historical datasets derived from clinical practice. The first option involves prospectively testing various dose concentrations on the same individual a closely matched group of individuals. However, intentionally deviating from the established clinical protocol to achieve varied datasets is not ethically feasible both in routine care. The second option remains difficult to define. While clinicians may adapt protocols when deemed necessary, they typically adhere to established guidelines, generating data with sparse diversity. This limitation creates a knowledge and intervention gap in addressing situations where drug dosing policies remain sub-optimal. However, the methodologies available to improve these policies are often not applicable due to the inherent nature of clinical practice and the historical datasets derived from it.

Our work strives to cover this gap, and the main contributions of the presented research are:

- A straightforward methodology (IDoser) to improve individualized drug dosing policies by using available observational datasets and field knowledge while simultaneously incorporating requirements of the specific problem.
- An application in the FSH dosing use case that shows significant improvements versus clinical practice and a literature benchmark.
- The validation, by achieving improvement versus clinical practice and the literature benchmark, that introducing field knowledge in the training process is key for the improvement of dosing policies.
- The possibility to customize the search for policy solutions depending on each specific dosing problem.
- A potential applicability for multiple typologies of dosing problems due to the flexible capabilities of our proposed approach.

2. Related work

The dose–response relationship describes an organism's response to a stimulus or stressor, typically a chemical or drug, in relation to the magnitude of exposure or dose, administered either once or over a period of time. Dose–response curves (Hayes et al., 2020) describe these relationships. A comprehensive understanding of this relationship is imperative for the development of dosing protocols and policies. Historically, curve functions have been employed to model this phenomenon (Calabrese, 2016; Sta et al., 2023).

Dose–response curves can exhibit diverse shapes, indicating the complex relationship between dose and response. One prominent theory underlying dose–response curves is hormesis. According to hormesis, substances may exhibit toxic effects at high doses, while lower doses can have beneficial or stimulatory effects on biological systems. Consequently, plotting the substance's benefit yields a U-shaped curve, whereby minimal doses lack efficacy, low to moderate doses exhibit a positive effect, and higher doses become increasingly detrimental (Abbaraju et al., 2023; Calabrese & Baldwin, 2002; Mohseni Ahooyi & Soroush, 2015).

Another important aspect of dose–response modeling is the threshold model, which suggests the existence of a threshold dose below which no clinically significant or detectable effect is observed. However, once the threshold is surpassed, the response increases proportionally to the dose, which can manifest as a linear dose–response function. Related to this model is the linear non-threshold model (LNT), commonly employed in radiation science and recently challenged, wherein even low doses are presumed to be harmful (Sacks et al., 2016; Selby & Calabrese, 2023).

The concept of saturation also plays a role in dose–response curves. Saturation occurs when a drug or substance achieves its maximum response at a specific dose, beyond which further dose increments do not elicit any additional effects. This saturation phenomenon can be represented by a sigmoidal dose–response curve, wherein the effect initially demonstrates rapid growth with increasing dose but eventually levels off as the substance reaches its maximum impact. Many biological response curves can be accurately approximated by a sigmoidal shape due to the involvement of saturation processes, such as the occupation of all available specific receptors for the substance. The Hill equation (Hill, 1910), a non-linear logistic function comprising four parameters, is frequently employed to fit these relationships (Gadagkar & Call, 2015).

In addition to saturation, sigmoid functions are also characterized by monotonicity. Positive monotonicity refers to the phenomenon where, as the dose increases, the corresponding outcome either increases or remains the same, but never decreases. This characteristic implies a consistent upward trend in the relationship between dose and response. FSH dose–response is no exception, as both literature and clinical experience indicate that a sigmoid function may be a good fit for its relationship with outcomes, including the number of retrieved oocytes. This is clearly reflected in the results of the pharmacometrics studies by Arce et al. (2016) and Porchet et al. (1994), where sigmoid functions of the type E-max were used to fit the PD portion and adequately described the study data. 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. Therefore, we infer that the dose–response relationship between FSH and the number of retrieved oocytes is at least positive and monotonic.

Model-Informed Precision Dosing or MIPD (Darwich et al., 2017; Del Valle-Moreno et al., 2023; Keizer et al., 2018; Poweleit et al., 2023) constitutes a good approach to individualize dose protocols, and its ideal form passes through obtaining well-fitted dose–response models based on PK/PD of the studied drug. Drugs approved for clinical use habitually have a published PK/PD model derived from phase III clinical trials.

Nevertheless, these models are often not applicable to all patients. Primarily, phase III PK/PD studies tend to exclude biomarkers known to affect the drug, although it is relatively common for individuals to have multiple comorbidities (Gonzalez et al., 2017). Second, trials often tend to include specific patient populations, excluding certain subgroups. For example, in the FSH dosing case, patients over 40 years of age and/or with irregular menstrual cycles, which are commonly encountered in IVF centers, have been previously excluded from studies (Barakhoeva et al., 2019; Bosch et al., 2019; Nyboe Andersen et al., 2017), even if these subpopulations are commonly found among IVF patients. Thirdly, assuming that the target population parameters have the same distributions as the study sample is often incorrect due to factors such as socioeconomic status, genetic and ethnic variations, and geographical differences (Keizer et al., 2018). Finally, PK/PD models may be fitted for outcomes that are not directly related to the clinical objective or do not consider key biomarkers known to affect

individual dose–response variability. This is the case for FSH dose in COS (Abd-Elaziz et al., 2017; Arce et al., 2016; Porchet et al., 1994).

When the developed model is not applicable to all subpopulations equally, the use of non-linear mixed methods (Del Valle-Moreno et al., 2023; Sheiner & Ludden, 1992; Sheiner & Steimer, 2000), physiologically based PK models (PBPK) (Jones et al., 2015), Bayesian methods (Darwich et al., 2017; Del Valle-Moreno et al., 2023; Hamberg et al., 2015; Sheiner & Beal, 1982), and PK/PD methods combined with machine learning (ML) (McComb & Ramanathan, 2020; Poweleit et al., 2023) can be applied to ameliorate the prediction. However, all of these approaches require an available covariate-linked PK/PD model as a starting point. While a PK/PD model can be developed, this approach is both computationally and labor-intensive and often requires data on drug blood concentrations after treatment (Koch et al., 2020; McComb et al., 2022). These requirements are often not met in clinical practice.

Whenever PK/PD methods are not applicable, ML approaches that rely on the concept of causal inference have been proposed. 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.

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 receiving a certain treatment) has been used to adjust for treatment selection bias. For multiple or continuous treatments, this concept is translated to the generalized propensity score (GPS) (Hirano & Imbens, 2005; Imbens, 2000). This score is used to weigh samples while estimating the outcome value. However, propensity score models must be precisely determined and can be numerically unstable due to extreme propensity weights (Bica et al., 2021; Peng et al., 2023).

Recent methods to ameliorate this problem include kernel functions to estimate the GPS (Colangelo & Lee, 2020; Kallus & Zhou, 2018) and Doubly Robust (DR) ML (Chernozhukov et al., 2018; Hoffmann, 2023) to estimate outcome values or use of bayesian procedures (Forastiere et al., 2022). Additionally, some studies employ the discretization of the treatment space (Cai et al., 2020; Schwab et al., 2019), or use generative adversarial methods (Bica & Jordon, 2020). While effective in estimating dose–response relationships, these approaches rely on two key assumptions that are essential for every causal inference analysis (Pearl et al., 2016):

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

In clinical settings, fulfilling these assumptions can be highly challenging. Clinicians largely adhere to clinically accepted dosing policies and tend to administer similar doses to patients with comparable characteristics. This practice limits the positivity assumption, as it frequently results in groups of patients with no available data in certain dose ranges. Furthermore, adjusting for all confounding variables uniformly across all cases may not always be possible. Clinicians may have personal and practice preferences for different biomarkers or may vary the extent of diagnostic tests based on factors such as expertise, financial considerations, or patient requests. Consequently, complying with the unconfoundedness assumption also constitutes a challenge.

In the specific case of selecting the starting FSH dose for COS, two studies have explored the use of nomograms (Ebid et al., 2021; La Marca et al., 2012), while Howles et al. (2006) applied a model based on multivariate regression to optimize individualized FSH dosing policies without relying on PK/PD or causal models. All three studies considered known relevant biomarkers, including patient age, anti-Müllerian hormone (AMH) levels, antral follicle count (AFC), and basal FSH levels. Subsequent RCTs evaluating the efficiency of these models showed promising results, including a reduction in the occurrence of OHSS (Olivennes et al., 2015), and an increase in the proportion of patients achieving optimal outcomes within the desired range of oocytes (Allegra et al., 2017). Nevertheless, these models were specifically developed for normo-ovulatory women under the age of 40 years, thereby excluding the most challenging patients in terms of dose selection.

Fanton et al. (2022) adopted a more comprehensive approach by computing individual FSH dose–response curves for all patient types. This was achieved by applying k-nearest neighbors (KNN) to identify the 100 most similar patients and fitting a constrained second-order polynomial to the data, specifically focusing on the number of mature oocytes retrieved and the administered FSH starting dose. Curves that were largely flat were categorized as non-responsive to FSH and accounted for 30% of the analyzed cases. For dose–response curves, the optimal starting dose was determined based on the presence of a peak in the mature oocyte curve. Using propensity score matching (PSM) to pair similar patients receiving different doses, the authors found that patients who received the optimal dose predicted by their model achieved better results compared to those who did not. While the study design is certainly interesting, the curves used to fit the dose–response relationship may not fit well with established pharmacometrics of FSH. Specifically, second-order polynomial curves do not adequately capture positive monotonicity, which is a key characteristic of the relationship between FSH and the number of mature oocytes retrieved.

3. Methods

3.1. The individualized dose improvement problem

To formalize the problem at hand, we introduced the Individualized Dosage Improvement Problem (IDIP). In this problem, we consider a large population of N patients denoted as $P = \{p_1, \dots, p_i, \dots, p_N\}$. The objective of the IDIP is to determine the optimal dosage of a given drug, which we refer to as the *dose*. A dose (d_i) and its corresponding outcome or *response* (y_i) are recorded for each patient p_i from population P. Each patient's response (y_i) is represented by a real number ($y_i \in \mathbb{R}$), while the dose (d_i) falls within the range $d_i \in [0, \infty)$. We assume that the response can be quantified by a single real number. Additionally, we assume that the desired levels of response (y_i^*) are known for each individual, indicating the target or optimal outcomes for the patients.

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 patient age, weight, height, and values of previous analysis, amongst others.

The *individualized dosage policy*, or IDP, $\pi : \mathcal{X} \to \mathbb{R}$ is a function that determines the *recommended dose* $\hat{d}_i$ based on the characteristics of the patient (x_i).

The objective was to find an IDP or π with the minimum error or *loss* possible, or π^*. Mathematically, this means identifying

$$\pi^* = \operatorname*{argmin}_{\pi} L(\pi), \tag{1}$$

where L is a *collective loss* function. Accordingly, the quality of the IDP π is inversely related to the value of $L(\pi)$. The collective loss $L(\pi)$ is computed as the average of *individual losses* (l_i), one for each patient p_i. The individual loss of a dose for a patient $l(y_i, y_i^*, \hat{d}_i, d_i)$ is a measure of how well a proposed dose ($\hat{d}_i$) would perform relative to the actual dose (d_i), its corresponding outcome y_i, and the desired objective (y_i^*). Mathematically, the *Collective loss* can be mathematically defined as

$$L(\pi) = \frac{\sum_{i=1}^{N} l(y_i, y_i^*, \pi(x_i), d_i)}{N}. \tag{2}$$

In the IDIP, we are given data that describes the current dosing practice for a particular drug. This data includes information for a subset of patients from a complete population, denoted as N. For each patient that has been administered the drug (p_i), where i ranges from 1 to N, we record

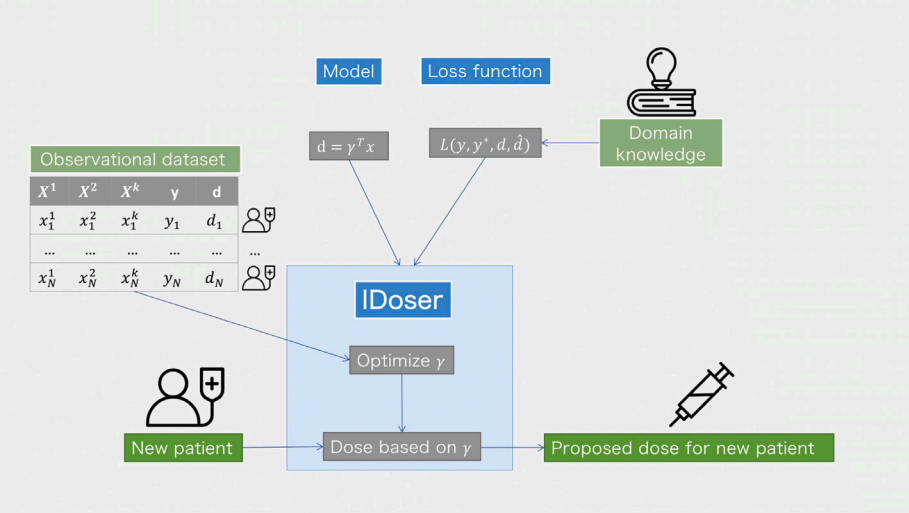


Fig. 1. Principal components of IDoser. A loss function, which integrates domain knowledge, is used to optimize γ of the selected core model using historical clinical data. With the optimized γ, new patients get personalized doses.

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

The main challenge in the IDIP is how best to leverage the information available from current dosing practices to develop a personalized dosing policy with minimal loss or error in dose selection.

3.2. Proposed approach: Individualized doser (IDoser)

Our proposed approach hinges on two key assumptions:

1. *Monotonicity* of the dose–response relationship. We assume that the dose–response relationship follows a monotonic relationship, meaning that as the dose increases, the expected response either increases or remains similar.
2. Knowledge of an *optimal outcome* (y^*). We assume that there is a known optimal outcome for the dosing problem, which can be specified as either a range or a point.

These two assumptions are grounded in the characteristics and goals of dosing problems. Monotonicity captures the key trend observed in threshold and linear non-threshold models, as well as sigmoid functions, which cover a wide range of dose–response relationships. On the other hand, knowledge of the optimal outcome aligns with the fundamental objective of dosing policies, which involves selecting the correct dose to achieve a desired outcome.

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

We will review both elements in the following subsections. Throughout the manuscript, positive monotonicity is assumed as default. Nevertheless, when required, IDoser can be readily adapted to accommodate a negative monotonicity assumption.

3.2.1. The core model

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

$$\hat{d}_i = \pi_\gamma(x_i), \tag{3}$$

where π_γ describes a core model or dosing policy, which is associated with a specific set of coefficients γ. The core model can be specified in its simplest linear form as

$$\hat{d}_i = \pi_\gamma(x_i) = \gamma^T \cdot x_i, \tag{4}$$

where γ^T denotes the transpose of the coefficient vector γ and $\cdot$ is the scalar product. This is the simplest form that satisfies the assumption of monotonicity and the requirement of relating the dose $\hat{d}_i$ to the covariates x_i through the coefficients γ. However, more complex forms can also be used as per the specific needs of the modeling approach.

3.2.2. Loss function

To achieve an improvement on any real dosing policy (π), it is crucial to be able to evaluate a hypothetical or counterfactual dose. To incorporate field knowledge into the evaluation, we encode key constraints knowledge into the loss function. Namely, we integrate our two assumptions: positive monotonicity and the existence of a desired outcome (y^*). This can be easily introduced as follows:

$$l(y, y^*, \hat{d}, d) = \begin{cases} -1 & \text{dose change is correct} \\ +1 & \text{dose change is incorrect} \\ 0 & \text{no dose change} \end{cases} \tag{5}$$

where a correct dose change is increasing d (the dose) whenever y (the obtained response) is less than y^* (the desired response), and decreasing d (the dose) when y (the obtained response) is greater than y^* (the desired response). Any change outside of these assumptions would be incorrect (Fig. 2). No change in dose, namely when $\hat{d} = d$ is considered neither good nor bad. Hence the loss value attributed is 0.

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

The function that generates the loss evaluation can be adapted as needed to accommodate negative monotonicity assumptions. Additional rules can be incorporated depending on the specific situation or requirements. This flexibility allows for the customization of the loss evaluation function to effectively handle various scenarios in different contexts.

One example is to define changes in the right direction (to the desired outcome) as beneficial while introducing limitations on the magnitude of those dose changes. This type of limitation would ensure that uncertainty is taken into account, as larger changes in dose imply less confidence in predicting its effect. This approach would allow for a

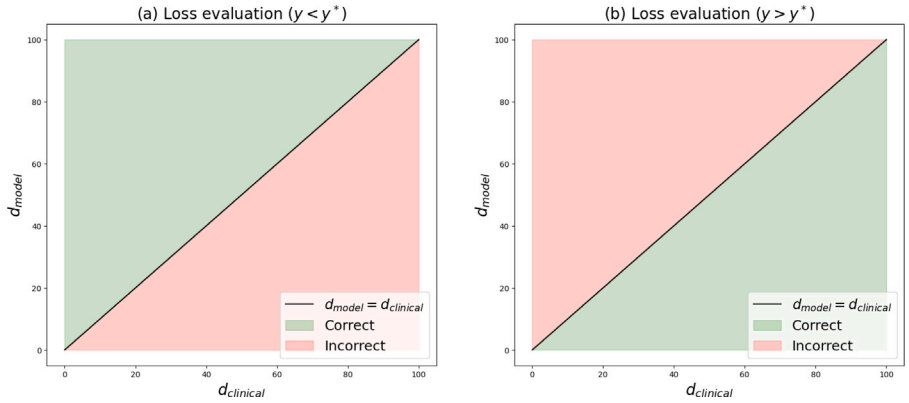


Fig. 2. Graphical representation of loss evaluation for cases (a) where y (the obtained response) is less than y^* (the optimal response), (b) where y (the obtained response) is greater than y^* (the optimal response).

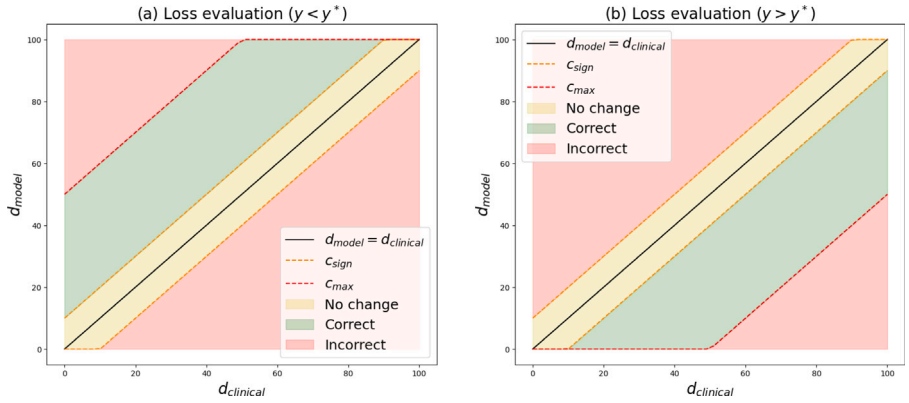


Fig. 3. Graphical representation of loss evaluation with additional rules considering a maximum change allowed (c_{max}) and a minimum change threshold (c_{sign}), (a) for cases where y (the obtained response) is less than y^* (the desired outcome) and (b) and where y (the obtained response) is greater than y^* (the desired outcome).

more nuanced assessment of dose adjustments, taking into account the level of confidence in predicting the response to dose changes.

Another rule that may be incorporated into the evaluation process involves introducing a threshold to determine when to consider a proposed dose ($\hat{d}$) different from the actual dose (d).

These additional rules would consequently impact the allocation of the loss value, as represented in Eq. (6) and in Fig. 3. We apply these rules in our dosing case, as shown below, to enhance the evaluation process and contribute to more informed and tailored dose decisions. The use of these additional rules will be specific to the use case and can enhance evaluation in different dosing scenarios.

$$l(y, y^*, \hat{d}, d) = \begin{cases} -1 & \text{dose change is correct and under a maximum change allowed } (c_{max}) \\ +1 & \text{dose change is incorrect or in correct direction but over } c_{max} \\ 0 & \text{no dose change or change under a minimum change considered significant } (c_{sign}) \end{cases} \tag{6}$$

3.2.3. Optimization of parameters

After defining the core model and the loss function according to the use case, the next step is to determine the parameters of the model by minimizing the loss function. Here, we propose the use of a coordinate descent algorithm (Wright, 2015) to iteratively establish the set of parameters that results in a minimum collective loss, denoted as L:

$$\gamma^* = \underset{\gamma}{\operatorname{argmin}} L(\pi_\gamma) = \frac{\sum_{i=1}^{N} l(y_i, y^*, \pi_\gamma(x_i), d_i)}{N} \tag{7}$$

Once the coordinate descent algorithm reaches a minimum, the resulting optimized parameters γ^* determine the optimized dosing policy, namely $\pi^* = \pi_{\gamma^*}$.

3.3. Use case

The use case applied in this study aims to determine the appropriate dose of FSH for COS during IVF treatment.

The available observational dataset consists of a set of covariates related to the ovarian reserve of patients, along with the dose of FSH prescribed by clinicians and the corresponding outcome, measured in the number of mature oocytes retrieved. The covariates captured in the dataset include patient age at the time of treatment, body mass index (BMI), AFC, AMH levels, and basal endogenous FSH levels. These factors are essential for assessing ovarian reserve and guiding the dose selection process. Two distinct databases were retrieved for this study. The first was dedicated to developing the dosing models and consists of first IVF cycle patients undergoing treatment between January 2011 and December 2019. The second database, separated temporally, was reserved only for validation of the resulting models and included cases from January 2020 to September 2021. The separation allowed an independent evaluation of the model's performance on unseen data. The development database was used to train and test the dosing models through 5-fold cross-validation in order to analyze the performance of multiple candidate models. Once analyzed and selected for validation, the final models were trained with the full development database. It is important to highlight that the validation dataset consists of newer cases and exhibits a slight shift in the clinical dosing policy compared to the development database. Specifically, the maximum dose of FSH in the validation dataset is considerably lower

Table 1
Summary of the characteristics and differences between the development and validation databases, expressed as mean ± standard deviation (SD) and range [minimum-maximum].

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

than the maximum value observed in the development database. This difference can be attributed to the gradual implementation of more conservative guidelines introduced by the European Society of Human Reproduction and Embryology (ESHRE) (The ESHRE Guideline Group on Ovarian Stimulation et al., 2020). Table 1 provides a summary of the characteristics of the development and validation databases.

Based on available literature, we confirmed that both assumptions required for our proposed approach are valid. While some evidence in cows (Karl et al., 2021) may challenge our first assumption, positive monotonicity, studies in human consistently demonstrate that an increase in FSH dose does not result in a decrease in the number of oocytes retrieved under the same conditions (same patient, same menstrual cycle) (Abd-Elaziz et al., 2017; Arce et al., 2016; Lensen et al., 2018; Porchet et al., 1994). In human, any negative effects of higher doses of FSH primarily relate to oocytes quality rather than quantity (Luo et al., 2022). Therefore, the positive monotonicity assumption holds true, indicating that increasing the FSH dose leads to an equal or greater number of retrieved oocytes. Nevertheless, oocyte quality should not be disregarded, as it is also a crucial factor for cycle success. Both quality and quantity of oocytes are important, as only retrieved and fertilized oocytes have the potential to develop into blastocysts, embryos with the highest chance of success (Maggiulli et al., 2020; Vaiarelli et al., 2020). Therefore, it is necessary to strike a balance, by defining an optimal number of oocytes for optimizing cycle outcomes. Regarding our second assumption, known optimal outcome, clinicians aim to select the first dose of FSH that will result in an optimal number of mature oocytes. However, there may be some variation in the specific definition of what constitutes an optimal number in literature (hui Chen et al., 2017; Ji et al., 2013; Polyzos & Sunkara, 2015; Steward et al., 2014; Sunkara et al., 2011). For our study, we defined optimal outcome between 10 (y^*_{min}) and 15 (y^*_{max}) mature oocytes, following the recommendations by Steward et al. (2014) and Sunkara et al. (2011). While the desired optimal outcome range applies to every patient, some patients with a reduced ovarian reserve may not be able to reach this range. For these patients, the dose will be adjusted as needed to bring them as close as possible to the optimal range. It should be noted that the dose (d_i) has a minimum value set by definition, which is at least 0 ($d_i \in [0, \infty)$). But this minimum, depending on the specific case can be higher than 0. Additionally, there is a maximum limitation on the dose, which can further impact the dosing options available to patients. In this specific use case, we set the minimum dose (d_{min}) at 100 IU of FSH, and explored doses ranging from 300 to 450 IU for the maximum dose (d_{max}).

3.3.1. IDoser for FSH dosing

The core model and the loss function are two essential elements for the application of our proposed IDoser. In this study, the selected core dosing model for FSH assumes a linear dose–response relationship and is defined as

$$y_i = y_0 + (\beta^T \cdot x_i)d, \tag{8}$$

where β is a set of coefficients related to a set of x_i of equal length, T represents the transposition of the vector β, $\cdot$ indicates the scalar product, and y_0 is the value of y when the dose (d) is 0. To derive the recommended dose ($\hat{d}_i$) based on a desired outcome (y^*), we rearrange Eq. (8) as follows:

$$\hat{d}_i = \frac{y^* - y_0}{\beta^T \cdot x_i}. \tag{9}$$

This can be further generalized to the following dosing model:

$$\hat{d}_i = \frac{\kappa}{\beta^T \cdot x_i}. \tag{10}$$

In this equation, κ is composed by the difference between y^* and y_0 component from Eq. (9). Parameters κ and β will form now γ, denoted as $\gamma = (\kappa, \beta)$. By incorporating parameters κ and β, the dosing model (Eq. (10)) allows the estimation of the optimal dose ($\hat{d}_i$) based on patient characteristics (x_i) and the desired outcome (y^*).

In addition to the basic rules described, the loss function incorporates additional rules to ensure an improved yet conservative dosing policy. These rules aim to discourage highly variable doses that may lead to greater uncertainty in achieving the expected outcome. These additional rules contribute to the conservative nature of the dosing policy. Limitations on dose changes were defined based on the desired outcome range for each specific patient. Following the definitions provided by Polyzos and Sunkara (2015), we established the following categories:

- 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;
- An outcome greater than 15 mature oocytes was considered too high.

Considering clinical implications and expert insights, the dosing policy incorporates different allowable dose modifications depending on the outcome range. Accordingly, patients with outcomes categorized as too-low or too-high are allowed higher dose modifications compared to those with sub-optimal outcomes. Specifically, dose modifications of up to 150 IU were allowed for patients with outcomes categorized as too-high or too-low, while a maximum dose modification of up to 75 IU was allowed for patients with sub-optimal outcomes. Changes up to 25 IU were not considered significant modifications. We established these thresholds based on the collective knowledge and expertise of multiple professionals in the field who regularly perform this task. Thresholds were determined through interviews and joint deliberation to ensure that clinical impacts were appropriately addressed.

Differential clinical impacts exist between outcomes that are too low and those that are too high. Too few mature oocytes can result in cycle cancellation, leading to no chance of pregnancy for patients, and higher economic and psychosocial burden. Conversely, an excessively high number of oocytes carry an increased risk of OHSS, which, if left untreated, can have detrimental consequences on the patient's health. Currently, there are well-established pharmacological interventions that all but eliminate the clinical manifestation of OHSS (Castillo et al., 2020; Najdecki et al., 2022). However, OHSS risk management entails the need to delay embryo transfer, which in turn results in a greater investment of time and money to achieve a safe pregnancy. Considering the detrimental consequences associated with both too-high and too-low outcomes, it is crucial to avoid these extremes in the dosing policy. By incorporating these considerations, the dosing policy aims to strike a balance between achieving an optimal outcome range whilst mitigating the risks and challenges associated with insufficient or excessive outcomes. The objective is to optimize the chances of successful treatment while prioritizing patient well-being and minimizing potential complications.

3.4. Evaluation methodology

3.4.1. Literature benchmark

We identified a relevant implementation in the study conducted by La Marca et al. (2012), which was later tested in an RCT (Allegra et al., 2017). This work uses a core model that shares similarities with our research, specifically ensuring positive monotonicity. Additionally, our second assumption, regarding a known optimal outcome, was addressed in the paper by fixing y^* to 9 oocytes for all patients, implicitly considering y_0 as 0. We have therefore chosen to use this implementation as the literature *benchmark* for our study, henceforth referring to it as the La Marca model or LM.

The covariates considered in the La Marca study included age, AMH, and basal FSH levels. The dosing model was derived by performing a linear regression of the following equation:

$$\frac{y_i}{d_i} = \beta^T \cdot x_i, \tag{11}$$

where the coefficients included in β were estimated to construct the dosing model and were made available in their study. The resulting dosing model constructed can be expressed as

$$\hat{d}_i = \frac{y^*}{\beta^T \cdot x_i}, \tag{12}$$

which is similar to our own proposed core model represented in Eq. (10).

In the subsequent RCT, Allegra et al. (2017) found that a significantly higher proportion of patients achieved an optimal outcome, defined as the retrieval of 8 to 14 oocytes, even though the mean number of oocytes did not show a significant change.

Using the published β values and Eq. (12), we established the LM model or policy, which we applied to the validation cases alongside our proposed approach. This allowed us to compare the performance of the LM model, an existing solution, with our novel approach.

3.4.2. Optimization exploration

We explored several approaches for optimizing the LM model, as described in Appendix A. These approaches were compared to both clinical practice and the unmodified LM model in Appendix B. The final model was obtained by incorporating two additional covariates from our dataset, namely AFC and BMI. The variable basal FSH was committed from our final model. All parameters of γ were optimized, leading to the development of the proposed IDoser.

3.4.3. Model comparison and statistic tests

We used two methods to compare the optimized models with the LM model and clinical practice. The first method involved analyzing and plotting collective loss L across all d_{max} within the allowed range. This first analysis provided insights into the quality of the IDoser and LM models. This extra comparison introduced an oracle decision policy or model that always makes the correct dose recommendations: if a dose change is required, it is always made in the right direction and within the adequate range. The oracle model represents the ideal policy according to our loss function as it captures all the correct dose changes for the test patients. 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 from each other, the loss values from each group were compared using the method recommended by García and Herrera (2008) and García et al. (2010), which is an extension of the study by Demšar (2006). Specifically, Iman–Davenport's corrected Friedmann test (Iman & Davenport, 1980) was employed, as the normality test did not meet the assumptions. When significant results were achieved, a post hoc test was conducted to determine which groups were different. The p-values were adjusted using Finner's correction, as per (García et al., 2010). The statistical methods were performed using the *scmamp* package in R. A p-value of less than 0.05 was considered significant.

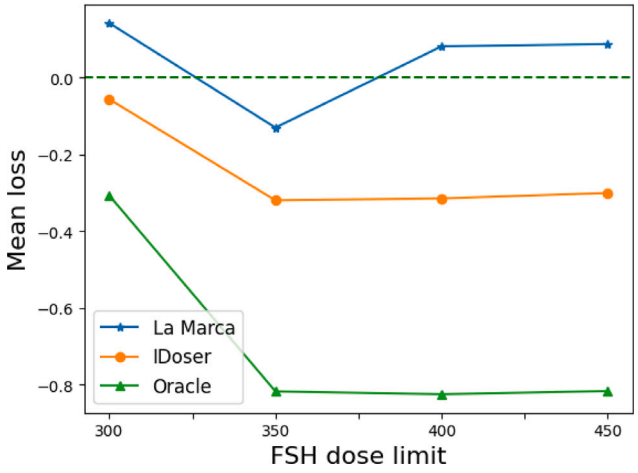


Fig. 4. Collective loss (L) across different d_{max} values for La Marca, the oracle model, and the proposed IDoser when applied to the validation dataset. The dashed line represents the L value for the clinical practice dosing method.

Table 2
Results of Iman–Davenport's corrected Friedman's rank sum test for all methods tested across the 4 selected values for d_{max}.

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

* P-values less than 0.05 were considered statistically significant.

4. Results

To compare our proposed model, IDoser, with the LM model and clinical practice we plotted the collective loss L for each model when applied to our validation dataset across all values of d_{max} explored. In the case of clinical practice, the L value was always 0. The obtained L values obtained were plotted in Fig. 4, alongside the L value of the oracle model, which always makes dose recommendations in the correct direction and within the appropriate range.

As seen in Fig. 4, compared to the LM model, IDoser consistently showed lower L values, also crossing 0, which represents the values for clinical practice. It is evident that the oracle model performs even better, as its L values were lower than both IDoser and the LM model. Given that the oracle model sets the benchmark for optimal performance, it is clear that there is still room for improvement in dosing policies.

The results of Iman–Davenport's corrected Friedman, as shown in Table 2, indicate a significant difference in loss values among the different models across all selected points of d_{max}.

Furthermore, post hoc test results, as shown in Table 3, demonstrate a significant improvement of our IDoser model compared to the LM model across all explored d_{max} points. This finding also holds true when comparing IDoser to clinical practice, except in the case of $d_{max} = 300$. Here, while an improvement was observed, it was not found to be significant. Specific loss values (L) and adjusted p-values are shown in Tables 4 to 7 in Appendix A.

Finally, given that each appropriate dose adjustment is designated an individual loss of −1, while any incorrect dose change is assigned a value of +1, and no modification receives a value of 0, the calculation of the collective loss L, when multiplied by 100, will yield the percentage of net error reduction. Conversely, when the sign is reversed, the resultant figure can be defined as the net improvement of the dosing

Table 3
Results of one vs. one comparison across all d_{max} values, ordered from worst (left) to best (right) method. These results were extracted from the post hoc test. P-values were adjusted by Finner's methodology.

d_{max}	Ordered results by significant differences
300	La Marca ≺ Clinical practice ∼ IDoser
350	Clinical practice ≺ La Marca ≺ IDoser
400	La Marca ∼ Clinical practice ≺ IDoser
450	

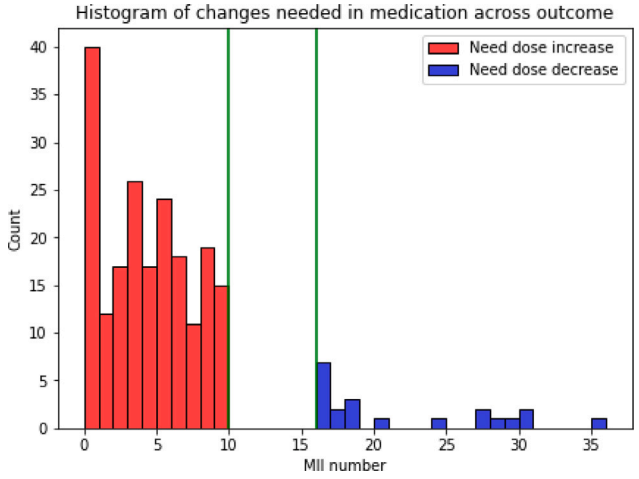


Fig. 5. Distribution of cases requiring an increase (red) or decrease (blue) in dose in the validation dataset with $d_{max} = 450$ is allowed.

policy in accordance with our defined loss rules. As such, it can be said that IDoser achieves a net improvement of 5.62% when $d_{max} = 300$, and up to 31.97% if $d_{max} = 350$.

5. Discussion

The proposed IDoser model demonstrated a significant improvement compared to the LM model across all investigated d_{max} values, as well as in the comparison to baseline clinical practice, with the exception of $d_{max} = 300$. Here, there is limited potential for further improvement in the dosing policy compared to other d_{max} values, as highlighted by Fig. 4. This is primarily attributed to the limited number of cases that could be improved and the inability of IDoser to identify a correct dose change for all of them. This is evident from the distribution of cases that can be improved, which were mainly concentrated in the low or suboptimal ranges of the outcome axis (number of mature oocytes), as shown in (Fig. 5).

These cases would require a significant increase in medication. However many low-responder patients have already received 300 IU of FSH from their clinicians, which is the most commonly used value for d_{max}, supported by the recent ESHRE guidelines for ovarian stimulation (The ESHRE Guideline Group on Ovarian Stimulation et al., 2020).

Fig. 6 illustrated that there are still some cases in the validation database that clinicians dosed over 300 IU. This suggests that clinicians ultimately prioritize individual patient characteristics rather than relying solely on population tendencies, especially if they believe that certain patients may benefit from exceeding the broadly recommended d_{max}. Given that the IDoser model has been optimized to automatically adhere to a more conservative dosing policy (with an optimized d_{max} of 333 IU), it can be used safely with an open d_{max} in order to identify

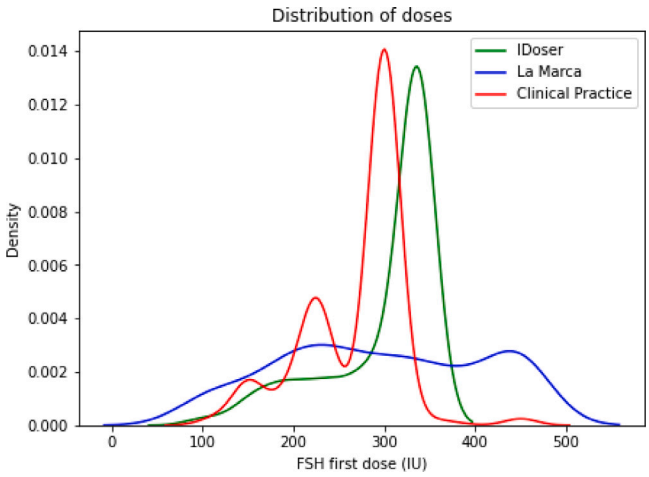


Fig. 6. Distribution of doses for the La Marca model (blue), clinical practice (red), and IDoser (green) in the validation dataset with $d_{max} = 450$.

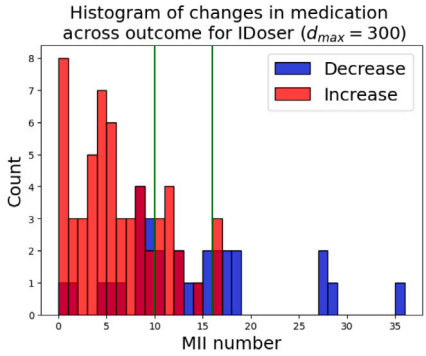
which patients that may benefit from an FSH dose over 300 IU of FSH. Fig. 6 provides a visual representation of how IDoser adjusts and shifts the dose distribution compared to clinical practice. In contrast, the LM model tends to distribute doses more evenly, with more cases receiving decreased doses and doses over 300 concentrated around the 450 IU mark (d_{max} value implemented in Fig. 6).

These antagonistic tendencies can also be observed in Figs. 7(a) to 7(d) that depict the distribution of dose changes distributed across outcome categories for both IDoser and the LM model with d_{max} of 300 and 450. IDoser (Figs. 7(a) and 7(c)) tends to rescue more cases that fall under our defined y^*_{min} range, where we find the majority of cases that can be improved. Notably, there are very few patients above y^*_{max} whose dose is increased and some patients whose dose is not decreased. This may be attributed to an under-representation of this subset of patients in our dataset and constitutes a limitation of the resulting model. On the other hand, with the LM model (Figs. 7(b) and 7(d)) more patients in need of a dose reduction receive it. However, many patients in need of a dose increase are instead given a reduced dose. Accordingly, our loss function ultimately penalizes the LM model.

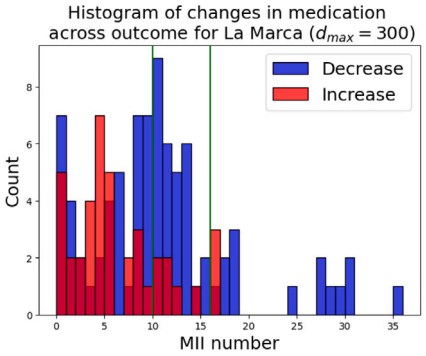
The decreasing tendency of the LM model may be attributed to the use of a y^* value of 9, which is lower than our defined y^*_{min} of 10. While this may explain the dose reductions for some patients in our database, it does not account for the relevant dose reductions observed in patients with very low outcomes (Figs. 7(b) and 7(d)). Another factor contributing to the lower performance of the LM model may be its limited scope, as it was originally developed for normo-ovulatory patients under 40 years, thus excluding a significant portion of the patient population who have different characteristics and ovarian reserve. In contrast, our database encompassed a wider range of patients, including all patients eligible for IVF treatment.

Also notable is the vantage obtained with IDoser while not using a single point for y^* but a range of desired outcomes. This makes a model less prone to change doses in the desired interval, as shown in Figs. 7(a) to 7(d).

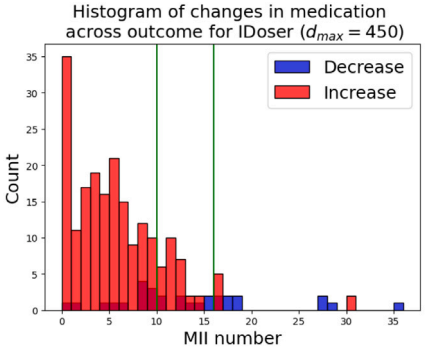
Finally, it is important to acknowledge that our loss function penalizes dose changes that are considered too large, even if they are made in the correct direction. Despite our efforts to optimize the model to avoid excessive changes, there are cases where significant changes in dose are still recommended. Yet, this may suggest that some patients genuinely require such dose modifications. Nevertheless, the true clinical utility of IDoser can only be demonstrated through a prospective RCT, which would offer further evidence to guide clinical decision-making and aid in validating the generalizability of IDoser across different patient populations and clinical settings.



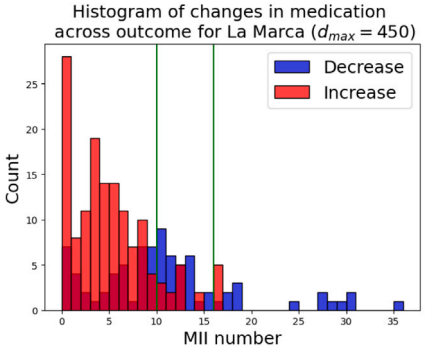
(a) Doses changes for the IDoser model with $d_{max} = 300$



(b) Doses changes for the La Marca model with $d_{max} = 300$



(c) Doses changes for the IDoser model with $d_{max} = 450$



(d) Doses changes for the La Marca model with $d_{max} = 450$

Fig. 7. Dose changes for (a) the IDoser model $d_{max} = 300$; (b) the La Marca model $d_{max} = 300$; (c) IDoser model with $d_{max} = 450$; and (d) the La Marca model with $d_{max} = 450$.

6. Conclusions

The implementation of the proposed approach, IDoser, in the FSH dosing case, demonstrated a significant improvement in dosing accuracy compared to clinical practice and a literature benchmark. By leveraging field knowledge and optimizing dosing policies, IDoser offers a practical solution to enhance patient care, that can be applied to similar dosing problems across different domains. The simplicity and effectiveness of IDoser also make it a valuable tool in situations where the available historical datasets are not amenable to more complex methodologies, such as causal inference and double machine learning.

Observational datasets available from clinical practice often have limited variability, which affects the generalizability and reliability of the algorithms that they are used to train. Clinical datasets may not capture all relevant confounding and often follow consistent biomarker selection protocols across different cases. This may be influenced by various factors, including the level of clinical experience, financial considerations, or even specific requests made by patients. Accordingly, algorithms trained solely on these datasets may not align with field knowledge or logical reasoning. Incorporating rules applied by clinicians based on their experience into a core model and a loss function is a valuable approach to developing dosing models. This approach allows for versatility and adaptability to different dosing settings, accommodating factors such as negative monotonicity or the use of different core functions. These concepts are partially inspired by PK/PD models, where physiological and pharmacological assumptions guide the modeling process. These principles are translated into our methodology by incorporating a core model that adheres to the monotonic assumption. Additionally, our approach includes specific rules within the loss function that penalize any dose change that deviates from the expected direction, maintaining the integrity of the dosing policy and optimizing the desired outcomes for patients.

While IDoser has proven effective in optimizing dosing policies, further improvements may provide more comprehensive dosing solutions. In this study, we explore a core function derived from a linear dose–response. In the future, other core models that align more closely with physiological dose–response relationships, such as sigmoid functions may be explored. Moreover, as IDoser is only applicable to single-dose dosing cases, further incorporation of dose adjustments over time for individual patients may be an important consideration for addressing different clinical scenarios. Finally, the implementation of individual values for the desired outcome (y^*) or the use of alternative optimization methodologies beyond coordinated descent can be explored, allowing for a more tailored and accurate approach to dosing.

Ultimately, IDoser achieved an optimized dosing policy in a time-efficient manner, but it can also be implemented conservatively 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 ensures a faster acceptance of the route of an RCT.

IDoser utilizes field knowledge and optimization techniques to improve dosing policies. Notably, IDoser may also be implemented in a conservative manner to validate drug doses through simulation (“in silico”). This is especially important, given that RCTs entail significant investments in time and resources. IDoser may thus provide a faster and more cost-effective strategy to evaluate the effectiveness and safety of dosing policies without the need for drug administration and patient involvement. Ultimately, the value of IDoser lies in its versatility, making it a valuable tool with the potential to enhance patient care, whilst also facilitating more informed discussions amongst stakeholders and supporting the decision-making process prior to embarking on RCTs.

Table 4
Posthoc test of differences in individual losses between explored methods and clinical baseline capping d_{max} at 300.

	Clinical practice	La Marca	IDoser κ	IDoser all variables	IDoser + AFC + BMI	IDoser without FSH
Collective loss	0	0.1423221	0.01872659	−0.04494382	−0.03745318	−0.05617978
Clinical practice						
La Marca	0.02998277*					
IDoser κ	0.85500958	0.043379244*				
IDoser all variables	0.54500150	0.004820011*	0.50871801			
IDoser+AFC+BMI	0.66092247	0.005758714*	0.55620182	0.818820510		
IDoser without FSH	0.54500150	0.004820011*	0.50871801	0.953885933	0.804125364	

* p-values adjusted using Finner method when under 0.05.

Table 5
Posthoc test of differences in individual losses between explored methods and clinical baseline capping d_{max} at 350.

	Clinical practice	La Marca	IDoser κ	IDoser all variables	IDoser + AFC + BMI	IDoser without FSH
Collective loss	0	−0.1301115	−0.2490706	−0.3048327	−0.2973978	−0.3197026
Clinical practice						
La Marca	3.402732e−02*					
IDoser κ	3.243974e−06*	0.017203405*				
IDoser all variables	9.522397e−08*	0.001441322*	5.058417e−01			
IDoser+AFC+BMI	3.703982e−07*	0.004658995*	7.258543e−01	7.258543e−01		
IDoser without FSH	9.522397e−08*	0.001441322*	5.058417e−01	9.724247e−01	7.258543e−01	

* p-values adjusted using Finner method when under 0.05.

CRediT authorship contribution statement

Nuria Correa: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. **Jesus Cerquides:** Conceptualization, Methodology, Software, Writing – review & editing, Supervision, Project Administration, Resources. **Rita Vassena:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Mina Popovic:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Josep Lluis Arcos:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project Administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

Permission to conduct this study was obtained from the Ethics Committee for Research with Medicinal Products, Eugin Barcelona (approval code: ALGO2).

We would like to thank Dr Daniel Mataró Marçal for generously dedicating his time and expertise on ovarian stimulation protocols that have greatly contributed to the development of IDoser and refined our research.

Funding

This work was supported by an industrial doctoral grant "Doctorat Industrial" funded by Ministry of Catalunya - Generalitat de Catalunya [DI-2019-24], by the project CI-SUSTAIN funded by the Spanish Ministry of Science and Innovation [PID2019-104156GB-I00], by the EU-ROVA Innovative Training Network (MSCA- ITN-2019-860960), and by intramural funding provided by Eugin Barcelona, part of the Eugin Group.

Nuria Correa is a PhD Student of the doctoral program in Computer Science at the Universitat Autonoma de Barcelona.

Appendix A. Optimization exploration

Given that by our definition y^*_{min} and y^*_{max} are slightly higher (10 to 15), we primarily used their β coefficients and optimized only κ. Next, our workflow was designed to explore the optimization of all coefficients in γ and the addition/omission of covariates. Specifically:

1. Optimization of only κ
2. Optimization of all γ
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 γ 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 up to a maximum of the same value. Then, every individual loss was evaluated. It is worth noting here that our models were limiting 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 doser.

Appendix B. Statistics results

See Tables 4–7.

Table 6
Posthoc test of differences in individual losses between explored methods and clinical baseline capping d_{max} at 400.

	Clinical practice	La Marca	IDoser κ	IDoser all variables	IDoser + AFC + BMI	IDoser without FSH
Collective loss	0	0.07777778	−0.2444444	−0.3	−0.2925926	−0.314814
Clinical practice						
La Marca	9.062553e−01					
IDoser κ	1.439933e−05*	8.548517e−06*				
IDoser all variables	5.412448e−07*	5.412448e−07*	5.366142e−01			
IDoser+AFC+BMI	1.790016e−06*	1.046312e−06*	7.413683e−01	7.582353e−01		
IDoser without FSH	5.412448e−07*	5.412448e−07*	5.366142e−01	9.816485e−01	7.582353e−01	

* p-values adjusted using Finner method when under 0.05.

Table 7
Posthoc test of differences in individual losses between explored methods and clinical baseline capping d_{max} at 450.

	Clinical practice	La Marca	IDoser κ	IDoser all variables	IDoser + AFC + BMI	IDoser without FSH
Collective loss	0	0.08791209	−0.2307692	−0.2857143	−0.2783883	−0.3003663
Clinical practice						
La Marca	7.585558e−01					
IDoser κ	3.540466e−05*	5.550225e−06*				
IDoser all variables	1.245251e−06*	3.574535e−07*	5.473513e−01			
IDoser+AFC+BMI	5.470478e−06*	1.245251e−06*	7.585558e−01	7.585558e−01		
IDoser without FSH	1.245251e−06*	3.574535e−07*	5.473513e−01	9.726274e−01	7.585558e−01	

* p-values adjusted using Finner method when under 0.05.

References

Abbaraju, V. D., Robinson, T. L., & Weiser, B. P. (2023). Modeling biphasic, non-sigmoidal dose–response relationships: Comparison of brain-cousens and cedergreen models for a biochemical dataset. ArXiv.

Abd-Elaziz, K., Duijkers, I., Stöckl, B., Klipping, C., Eckert, K., & Goletz, S. (2017). A new fully human recombinant FSH (follitropin epsilon): Two phase i randomized placebo and comparator-controlled pharmacokinetic and pharmacodynamic trials. *Human Reproduction*, *32*, 1639–1647. http://dx.doi.org/10.1093/humrep/dex220.

Allegra, A., Marino, A., Volpes, A., Coffaro, F., Scaglione, P., Gullo, S., & La Marca, A. (2017). A randomized controlled trial investigating the use of a predictive nomogram for the selection of the FSH starting dose in IVF/ICSI cycles. *Reproductive BioMedicine Online*, *34*, 429–438. http://dx.doi.org/10.1016/j.rbmo.2017.01.012.

Arce, J. C., Klein, B. M., & Erichsen, L. (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).

Barakhoeva, Z., Vovk, L., Fetisova, Y., Marilova, N., Ovchinnikova, M., Tischenko, M., Scherbatyuk, Y., Kolotovkina, A., Miskun, A., Kasyanova, G., Teterina, T., Zorina, I., Belousova, N., Morozova, E., Yakovenko, S., Apryshko, V., Sichinava, L., Shalina, R., & Polzikov, M. (2019). A multicenter, randomized, phase III study comparing the efficacy and safety of follitropin alpha biosimilar and the original follitropin alpha. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, *241*, 6–12. http://dx.doi.org/10.1016/j.ejogrb.2019.07.032.

Bica, I., Alaa, A. M., Lambert, C., & van der Schaar, M. (2021). From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology and Therapeutics*, *109*, 87–100. http://dx.doi.org/10.1002/cpt.1907.

Bica, I., & Jordon, J. (2020). Estimating the effects of continuous-valued interventions using generative adversarial networks. arXiv, 2002.12326v2.

Bosch, E., Havelock, J., Martin, F. S., Rasmussen, B. B., Klein, B. M., Mannaerts, B., & Arce, J. C. (2019). Follitropin delta in repeated ovarian stimulation for IVF: a controlled, assessor-blind phase 3 safety trial. *Reproductive BioMedicine Online*, *38*, 195–205. http://dx.doi.org/10.1016/j.rbmo.2018.10.012.

Cai, H., Shi, C., Song, R., & Lu, W. (2020). Deep jump Q-evaluation for offline policy evaluation in continuous action space. arXiv, 2010.15963.

Calabrese, E. J. (2016). The emergence of the dose–response concept in biology and medicine. *International Journal of Molecular Sciences*, *17*, http://dx.doi.org/10.3390/ijms17122034.

Calabrese, E. J., & Baldwin, L. A. (2002). Defining hormesis. *Human and Experimental Toxicology*, *21*, 91–97. http://dx.doi.org/10.1191/0960327102ht217.

Castillo, J. C., Haahr, T., Martínez-Moya, M., & Humaidan, P. (2020). Gonadotropin-releasing hormone agonist for ovulation trigger–OHSS prevention and use of modified luteal phase support for fresh embryo transfer. *Upsala Journal of Medical Sciences*, *125*, 131–137. http://dx.doi.org/10.1080/03009734.2020.1736696.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*, C1–C68. http://dx.doi.org/10.1111/ectj.12097.

Colangelo, K., & Lee, Y.-Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments. (pp. 1–39). arXiv, 2004.03036.

Darwich, A. S., Ogungbenro, K., Vinks, A. A., Powell, J. R., Marsousi, N., Daali, Y., Fairman, D., Cook, J., & Lesko, L. J. (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.. *Clin Pharmacol Ther.*, *101*, 646–656.

Del Valle-Moreno, P., Suarez-Casillas, P., Mejías-Trueba, M., Ciudad-Gutiérrez, P., Guisado-Gil, A. B., Gil-Navarro, M. V., & Herrera-Hidalgo, L. (2023). Model-informed precision dosing software tools for dosage regimen individualization: A scoping review. *Pharmaceutics*, *15*(1859), http://dx.doi.org/10.3390/pharmaceutics15071859.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Ebid, A. H. I., Motaleb, S. M., Mostafa, M. I., & Soliman, M. M. (2021). Novel nomogram-based integrated gonadotropin therapy individualization in vitro fertilization/ intracytoplasmic sperm injection: A modeling approach. *Clinical and Experimental Reproductive Medicine*, *48*, 163–173. http://dx.doi.org/10.5653/cerm.2020.03909.

Fanton, M., Nutting, V., Rothman, A., Maeder-York, P., Hariton, E., Barash, O., Weckstein, L., Sakkas, D., Copperman, A. B., & Loewke, K. (2022). An interpretable machine learning model for individualized gonadotrophin starting dose selection during ovarian stimulation. *Reproductive BioMedicine Online*, *45*, 1152–1159. http://dx.doi.org/10.1016/j.rbmo.2022.07.010.

Forastiere, L., Mealli, F., Wu, A., & Airoldi, E. M. (2022). Estimating causal effects under network interference with Bayesian generalized propensity scores. *Journal of Machine Learning Research*, *23*, 1–61.

Gadagkar, S. R., & Call, G. B. (2015). Computational tools for fitting the hill equation to dose–response curves. *Journal of Pharmacological and Toxicological Methods*, *71*, 68–76. http://dx.doi.org/10.1016/j.vascn.2014.08.006.

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180*, 2044–2064. http://dx.doi.org/10.1016/j.ins.2009.12.010.

García, S., & Herrera, F. (2008). An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research*, *9*, 2677–2694.

Gonzalez, D., Rao, G. G., Bailey, S. C., Brouwer, K. L., Cao, Y., Crona, D. J., Kashuba, A. D., Lee, C. R., Morbitzer, K., Patterson, J. H., Wiltshire, T., Easter, J., Savage, S. W., & Powell, J. R. (2017). Precision dosing: Public health need, proposed framework, and anticipated impact. *Clinical and Translational Science*, *10*, 443–454. http://dx.doi.org/10.1111/cts.12490.

Hamberg, A. K., Hellman, J., Dahlberg, J., Jonsson, E. N., & Wadelius, M. (2015). A Bayesian decision support tool for efficient dose individualization of warfarin in adults and children. *BMC Medical Informatics and Decision Making*, *15*, 1–9. http://dx.doi.org/10.1186/s12911-014-0128-0.

Hayes, A. W., Wang, T., & Dixon, D. (2020). Chapter 2 - dose and dose–response relationships in toxicology. In A. W. Hayes, T. Wang, & D. Dixon (Eds.), *Loomis's essentials of toxicology (Fifth Edition)* (Fifth). (pp. 17–31). Academic Press., http://dx.doi.org/10.1016/B978-0-12-815921-7.00002-8.

Hill, A. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves proc. *Physiol. Soc.*, *40*.

Hirano, K., & Imbens, G. W. (2005). The propensity score with continuous treatments. In *Applied bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with donald rubin's statistical family* (pp. 73–84). http://dx.doi.org/10.1002/0470090456.ch7.

Hoffmann, N. I. (2023). Double robust, flexible adjustment methods for causal inference: an overview and an evaluation. *American sociological association annual meeting 2024*.

Howles, C. M., Saunders, H., Alam, V., & Engrand, P. (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. *Current Medical Research and Opinion*, *22*, 907–918. http://dx.doi.org/10.1185/030079906X104678.

hui Chen, Y., Wang, Q., nan Zhang, Y., Han, X., han Li, D., & lian Zhang, C. (2017). Cumulative live birth and surplus embryo incidence after frozen-thaw cycles in PCOS: how many oocytes do we need? *Journal of Assisted Reproduction and Genetics*, *34*, 1153–1159. http://dx.doi.org/10.1007/s10815-017-0959-6.

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of the friedman statistic. *Communications in Statistics. Theory and Methods*, *9*, 571–595. http://dx.doi.org/10.1080/03610928008827904.

Imbens, G. (2000). The role of propensity score in estimating dose–response functions. *Biometrika*, 706–710.

Ji, J., Liu, Y., Tong, X. H., Luo, L., Ma, J., & Chen, Z. (2013). The optimum number of oocytes in IVF treatment: An analysis of 2455 cycles in China. *Human Reproduction*, *28*, 2728–2734. http://dx.doi.org/10.1093/humrep/det303.

Jones, H., Chen, Y., Gibson, C., Heimbach, T., Parrott, N., Peters, S., Snoeys, J., Upreti, V., Zheng, M., & Hall, S. (2015). Physiologically based pharmacokinetic modelling in drug discovery and development: A pharmaceutical industry perspective. *Clinical Pharmacology and Therapeutics*, 247–262.

Kallus, N., & Zhou, A. (2018). Policy evaluation and optimization with continuous treatments. *vol. 7*, In *International conference on artificial intelligence and statistics, AISTATS 2018* (pp. 1243–1251).

Karl, K. R., Jimenez-Krassel, F., Gibbings, E., Ireland, J. L., Clark, Z. L., Tempelman, R. J., Latham, K. E., & Ireland, J. J. (2021). Negative impact of high doses of follicle-stimulating hormone during superovulation on the ovulatory follicle function in small ovarian reserve dairy heifers. *Biology of Reproduction*, *104*, 695–705. http://dx.doi.org/10.1093/biolre/ioaa210.

Keizer, R. J., Heine, R. ter., Frymoyer, A., Lesko, L. J., Mangat, R., & Goswami, S. (2018). Model-informed precision dosing at the bedside: Scientific challenges and opportunities. *CPT: Pharmacometrics and Systems Pharmacology*, *7*, 785–787. http://dx.doi.org/10.1002/psp4.12353.

Koch, G., Pfister, M., Daunhawer, I., Wilbaux, M., Wellmann, S., & Vogt, J. E. (2020). Pharmacometrics and machine learning partner to advance clinical data analysis. *Clinical Pharmacology and Therapeutics*, *107*, 926–933. http://dx.doi.org/10.1002/cpt.1774.

La Marca, E., Grisendi, V., Argento, C., Giulini, S., & Volpe, A. (2012). Development of a nomogram based on markers of ovarian reserve for the individualisation of the follicle-stimulating hormone starting dose in vitro fertilisation cycles. *BJOG: An International Journal of Obstetrics and Gynaecology*, *119*, 1171–1179. http://dx.doi.org/10.1111/j.1471-0528.2012.03412.x.

Lensen, S. F., Wilkinson, J., Leijdekkers, J. A., Marca, A. La., Mol, B. W. J., Marjoribanks, J., Torrance, H., & Broekmans, F. J. (2018). Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). *Cochrane Database of Systematic Reviews*, *2018*, http://dx.doi.org/10.1002/14651858.CD012693.pub2.

Luo, X., Pei, L., He, Y., Li, F., Han, W., Xiong, S., Han, S., Li, J., Zhang, X., Huang, G., & Ye, H. (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. *Frontiers in Endocrinology*, *13*, 1–10. http://dx.doi.org/10.3389/fendo.2022.986438.

Maggiulli, R., Cimadomo, D., Fabozzi, G., Papini, L., Dovere, L., Ubaldi, F. M., & Rienzi, L. (2020). The effect of ICSI-related procedural timings and operators on the outcome. *Human Reproduction*, *35*, 32–43. http://dx.doi.org/10.1093/humrep/dez234.

McComb, M., Bies, R., & Ramanathan, M. (2022). Machine learning in pharmacometrics: Opportunities and challenges. *British Journal of Clinical Pharmacology*, *88*, 1482–1499. http://dx.doi.org/10.1111/bcp.14801.

McComb, M., & Ramanathan, M. (2020). Generalized pharmacometric modeling, a novel paradigm for integrating machine learning algorithms: A case study of metabolomic biomarkers. *Clinical Pharmacology and Therapeutics*, *107*, 1343–1351. http://dx.doi.org/10.1002/cpt.1746.

Mohseni Ahooyi, J. E., & Soroush, M. (2015). An efficient copula-based method of identifying regression models of non-monotonic relationships in processing plants. *Chemical Engineering Science*, *136*, 106–114. http://dx.doi.org/10.1016/j.ces.2015.03.044, URL https://www.sciencedirect.com/science/article/pii/S0009250915002250. Control and Optimization of Smart Plant Operations.

Najdecki, R., Michos, G., Peitsidis, N., Timotheou, E., Chartomatsidou, T., Kakanis, S., Chouliara, F., Mamopoulos, A., & Papanikolaou, E. (2022). Agonist triggering in oocyte donation programs—Mini review. *Frontiers in Endocrinology*, *13*, 1–7. http://dx.doi.org/10.3389/fendo.2022.838236.

Nyboe Andersen, A., Nelson, S. M., Fauser, B. C., García-Velasco, J. A., Klein, B. M., Arce, J. C., Tournaye, H., De Sutter, P., Decleer, W., Petracco, A., Borges, E., Barbosa, C. P., Havelock, J., Claman, P., Yuzpe, A., Višnová, H., Ventruba, P., Uher, P., Mrazek, M., ... Arce, J. C. (2017). Individualized versus conventional ovarian stimulation for in vitro fertilization: a multicenter, randomized, controlled, assessor-blinded, phase 3 noninferiority trial. *Fertility and Sterility*, *107*, 387–396. http://dx.doi.org/10.1016/j.fertnstert.2016.10.033, e4.

Olivennes, F., Trew, G., Borini, A., Broekmans, F., Arriagada, P., Warne, D. W., & Howles, C. M. (2015). Randomized, controlled, open-label, non-inferiority study of the CONSORT algorithm for individualized dosing of follitropin alfa. *Reproductive BioMedicine Online*, *30*, 248–257. http://dx.doi.org/10.1016/j.rbmo.2014.11.013.

Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics*, *6*, Article 7.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Peng, J., Zou, H., Liu, J., Li, S., Jiang, Y., Pei, J., & Cui, P. (2023). Offline policy evaluation in large action spaces via outcome-oriented action grouping. In *ACM web conference 2023 - Proceedings of the world wide web conference* (pp. 1220–1230). WWW 2023, http://dx.doi.org/10.1145/3543507.3583448.

Polyzos, N. P., & Sunkara, S. K. (2015). Sub-optimal responders following controlled ovarian stimulation: An overlooked group? *Human Reproduction*, *30*, 2005–2008. http://dx.doi.org/10.1093/humrep/dev149.

Porchet, H. C., Le Cotonnec, J. Y., & Loumaye, E. (1994). Clinical pharmacology of recombinant human follicle-stimulating hormone. III. Pharmacokinetic-pharmacodynamic modeling after repeated subcutaneous administration. *Fertility and Sterility*, *61*, 687–695. http://dx.doi.org/10.1016/s0015-0282(16)56646-1.

Poweleit, E. A., Vinks, A. A., & Mizuno, T. (2023). Artificial intelligence and machine learning approaches to facilitate therapeutic drug management and model-informed precision dosing. *Therapeutic Drug Monitoring*, *45*, 143–150. http://dx.doi.org/10.1097/FTD.0000000000001078.

Sacks, B., Meyerson, G., & Siegel, J. A. (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 authors). *Biological Theory*, *11*, 69–101. http://dx.doi.org/10.1007/s13752-016-0244-4.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., & Karlen, W. (2019). Learning counterfactual representations for estimating individual dose-response curves. arXiv, arXiv:1902.00981v3.

Selby, P. B., & Calabrese, E. J. (2023). How self-interest and deception led to the adoption of the linear non-threshold dose response (lnt) model for cancer risk assessment. *Science of the Total Environment*, *898*, Article 165402. http://dx.doi.org/10.1016/j.scitotenv.2023.165402, URL https://www.sciencedirect.com/science/article/pii/S0048969723040251.

Sheiner, L. B., & Beal, S. L. (1982). Bayesian individualization of pharmacokinetics: Simple implementation and comparison with non-Bayesian methods. *Journal of Pharmaceutical Sciences*, *71*, 1344–1348. http://dx.doi.org/10.1002/jps.2600711209.

Sheiner, L. B., & Ludden, T. M. (1992). Population pharmacokinetics/dynamics. *Annual Review of Pharmacology and Toxicology*, *32*, 185–209. http://dx.doi.org/10.1146/annurev.pa.32.040192.001153.

Sheiner, L. B., & Steimer, J.-L. (2000). Pharmacokinetic/pharmacodynamic modeling in drug development. *Annual Review of Pharmacology and Toxicology*, *40*, 67–95. http://dx.doi.org/10.1146/annurev.pharmtox.40.1.67.

Sta, L., Adamer, M. F., & Molina-París, C. (2023). Algebraic study of receptor–ligand systems: A dose–response analysis. *SIAM Journal of Applied Mathematics*, S105–S150. http://dx.doi.org/10.1137/22M1506262.

Steward, R. G., Lan, L., Shah, A. A., Yeh, J. S., Price, T. M., Goldfarb, J. M., & Muasher, S. J. (2014). Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: An analysis of 256, 381 in vitro fertilization cycles. *Fertility and Sterility*, *101*, 967–973. http://dx.doi.org/10.1016/j.fertnstert.2013.12.026.

Sunkara, S. K., Rittenberg, V., Raine-Fenning, N., Bhattacharya, S., Zamora, J., & Coomarasamy, A. (2011). Association between the number of eggs and live birth in IVF treatment: An analysis of 400 135 treatment cycles. *Human Reproduction*, *26*, 1768–1774. http://dx.doi.org/10.1093/humrep/der106.

The ESHRE Guideline Group on Ovarian Stimulation, Bosch, E., Broer, S., Griesinger, G., Grynberg, M., Humaidan, P., Kolibianakis, E., Kunicki, M., Marca, A. La., Lainas, G., Clef, N. Le., Massin, N., Mastenbroek, S., Polyzos, N., Sunkara, S. K., Timeva, T., Töyli, J., Vermeulen, N., & Broekmans, F. (2020). ESHRE guideline: ovarian stimulation for IVF/icsi†. *Human Reproduction Open*, *2020*, http://dx.doi.org/10.1093/hropen/hoaa009.

Vaiarelli, A., Cimadomo, D., Conforti, A., Schimberni, M., Giuliani, M., D'Alessandro, P., Colamaria, S., Alviggi, C., Rienzi, L., & Ubaldi, F. M. (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. *Fertility and Sterility*, *113*, 121–130. http://dx.doi.org/10.1016/j.fertnstert.2019.09.012.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, *151*, 3–34.